Decentralized but Globally Coordinated Biodiversity Data

Beckett W. Sterner[1*], Edward E. Gilbert[1], Nico M. Franz[1]

[1]School of Life Sciences, Arizona State University, United States

Submitted to Journal:
Frontiers in Big Data

Specialty Section:
Data Mining and Management

Article type:
Original Research Article

Manuscript ID:
509976

Received on:
04 Nov 2019

Frontiers website link:
www.frontiersin.org

Conflict of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

Author contribution statement

Sterner led development of the paper, with primary contributions to theoretical framing of decentralized coordination model (Intro, Sections 2 and 4, Conclusion). Gilbert's primary contributions were to work on the design, description, and implementation of decentralized coordination model (Section 5). Franz contributed to the overall message/conceptualization of the paper, especially to critique of current model and implementation of alternative for biodiversity data (Sections 3 and 5).

Keywords

Data aggregation (DA), ontology alignment (OA), Biodiversity data accessibility and use, Communities of Practice (CoP), Data intelligence, Decentralization, Knowledge commons, systematics

Abstract

Word count: 236

Centralized biodiversity data aggregation is too often failing societal needs due to pervasive and systemic data quality deficiences. We argue for a novel philosophy for biodiversity data science that embodies the spirit of the Web ("small pieces loosely joined") through the decentralized coordination of data across scientific languages and communities. Our decentralized approach encourages the emergence of multiple self-identifying communities of practice that are regionally, taxonomically, or institutionally localized. Each community is empowered to control the social and informational design and versioning of their local data infrastructures and signals. We argue that the upfront cost of decentralization is more than offset by the long-term benefit of achieving sustained expert engagement, higher-quality data products, and ultimately more societal impact for biodiversity data. With no single aggregator to exert centralized control over biodiversity data, decentralization generates loosely connected networks of mid-level aggregators. Global coordination is nevertheless feasible through automatable data sharing agreements that enable efficient propagation and translation of biodiversity data across communities. In other words, the global effort of coordination should focus on developing shared languages for data signal translation, as opposed to homogenizing the data signal itself. This decentralized model poses novel integration challenges, among which the explicit and continuous articulation of conflicting systematic classifications and phylogenies remain the most challenging. We present a critical review of the development of available solutions, challenges, and next pragmatic steps towards realizing decentralized, yet globally coordinated, biodiversity data services.

Contribution to the field

We argue for a novel philosophy for biodiversity data science that embodies the spirit of the Web ("small pieces loosely joined") through the decentralized coordination of data across scientific languages and communities. Our decentralized approach encourages the emergence of multiple self-identifying communities of practice who create, use, and curate biodiversity datasets that are regionally, taxonomically, or institutionally localized. Each community is empowered to control the social and informational design and versioning of their local data infrastructures and signals. We argue that the upfront cost of decentralization is more than offset by the long-term benefit of achieving sustained expert engagement, higher-quality data products, and ultimately more societal impact for biodiversity data. We present a critical review of the development of available solutions, challenges, and next pragmatic steps towards realizing decentralized, yet globally coordinated, biodiversity data services.

Funding statement

The Marine Biological Laboratory supported a 2017 workshop for concept development via funding from the McDonnell Foundation. Additional support for design conceptualization and implementation of data aggregation and translation tools includes Strategic Initiative Funding for the Biodiversity Data Science Initiative at Arizona State University and National Science Foundation grants DEB 1754731, STS 1827993, and DBI 1724433.

Decentralized but Globally Coordinated Biodiversity Data

Authors
Beckett W. Sterner*, Edward E. Gilbert, Nico M. Franz
School of Life Sciences
Arizona State University
*Corresponding author: beckett.sterner@asu.edu

Abstract: Centralized biodiversity data aggregation is too often failing societal needs due to pervasive and systemic data quality deficiencies. We argue for a novel philosophy for biodiversity data science that embodies the spirit of the Web ("small pieces loosely joined") through the decentralized coordination of data across scientific languages and communities. Our decentralized approach encourages the emergence of multiple self-identifying communities of practice that are regionally, taxonomically, or institutionally localized. Each community is empowered to control the social and informational design and versioning of their local data infrastructures and signals. We argue that the upfront cost of decentralization is more than offset by the long-term benefit of achieving sustained expert engagement, higher-quality data products, and ultimately more societal impact for biodiversity data. With no single aggregator to exert centralized control over biodiversity data, decentralization generates loosely connected networks of mid-level aggregators. Global coordination is nevertheless feasible through automatable data sharing agreements that enable efficient propagation and translation of biodiversity data across communities. In other words, the global effort of coordination should focus on developing shared languages for data signal translation, as opposed to homogenizing the data signal itself. This decentralized model poses novel integration challenges, among which the explicit and continuous articulation of conflicting systematic classifications and phylogenies remain the most challenging. We present a critical review of the development of available solutions, challenges, and next pragmatic steps towards realizing decentralized, yet globally coordinated, biodiversity data services.

Keywords: Data aggregation, ontology alignment, biodiversity data, communities of practice, data intelligence, decentralization, knowledge commons, systematics

Word count: 10,791
Number of figures: 6

1. Introduction

Open science policies are driving the emergence of a new, digital biodiversity knowledge commons. Inspired by big science efforts like the Human Genome Project, the goal has been to aggregate all data about where and when different biological entities — most typically "species" for our context — are located, in order to provide critical insight into global problems such as rapid biodiversity loss and climate change (Peterson et al. 2010, Devictor and Bensaude-Vincent 2016). An exceptionally large and heterogeneous set of stakeholders have a say in this emerging global biodiversity knowledge commons, making effective governance a major and ongoing challenge (Alphandéry and Fortier 2010, Turnhout et al. 2014). Historically, the dominant strategy has been to develop large initiatives aiming to *centralize* access through national or global web portals; using a unified metadata system to build a comprehensive, aggregated database and related services (Wieczorek et al. 2012). After several decades of these efforts to build a centralized and unified global infrastructure, the broader community is now in a reflective moment about where it should go next and what it can learn from past successes and failures. Widespread deficiencies in data quality and ultimately limited impact in relation to the societal investment, are fueling this dynamic of dissatisfaction (Mesibov 2013, 2018, Franz and Sterner 2018). This marks a pivotal moment for assessing how an alternative, *decentralized* approach can provide the "flexibility both to accommodate and to benefit from this diversity [of contributors], rather than seeking to implement a prescriptive programme of planned deliverables" (Hobern et al. 2019, 9) — as recommended by a recent report from the second Global Biodiversity Informatics Conference.

In contrast, leading paradigms for data discovery and integration in biology have presupposed stable consensus among multiple scientific communities about the best way to classify natural phenomena (Godfray 2002; Smith et al. 2007; de Jong et al. 2015; Ruggiero et al. 2015; Sterner et al. In Press). Following this model, first-generation big biodiversity data projects generally prioritized scale and comprehensiveness, though at the price of imposing uniformity on data that were generated using different standards and classification schemes.

The primary unit of data being aggregated is an observation event of an organism at a particular place and time, vouchered with a preserved material specimen, photograph, or in some cases DNA sequence. Historically, the dominant source for such "occurrence data" — as named by the Darwin Core standard (Wieczorek et al. 2012) — have been natural history collections and ecological surveys hosted and carried out around the world. The past several decades have introduced major new data sources, though, including enthusiast science efforts such as eBird or iNaturalist, remote sensing technologies such as camera traps or satellite imaging, and automated genetic sequencing efforts such as DNA barcoding and environmental metagenomics. With so many diverse projects and data sources, some sort of regional to global coordination to is clearly necessary (Turnhout and Boonman-Berson 2011). But this should also raise questions about the optimality of centralizing all data into one place and under a unified metadata system.

Next-generation solutions, we argue, must now reckon with the limitations of overriding local differences in the production and meaningfulness of biodiversity data (Franz and Sterner 2018). Not only is the fundamental nature of biodiversity highly contested, our knowledge of many taxonomic groups is vastly incomplete, changing at an unprecedented rate, and upending historically established methodologies (Hinchliff et al. 2015; Hortal et al. 2015; Whitman 2015; Peterson and Soberón 2017). More broadly, biodiversity loss is a good example of a "wicked" problem: "one cannot understand the problem without knowing about its context; one cannot meaningfully search for information without the orientation of a solution concept; one cannot first understand, then solve" (Rittel and Weber 1973: 162). Incorporating the possibility of dissent into the fundamental architecture of the data ecosystem may therefore prove crucial to harnessing the data revolution to address biodiversity loss.

Responsibility for persistent dissatisfaction with aggregated biodiversity cannot be simply or even primarily assigned to the globally distributed set of organizations holding biodiversity data collections (Franz and Sterner 2018). Many deficiencies occur *through* the aggregation process, i.e., at higher levels in the data flow trajectory. Moving beyond diagnosing the issues, we now turn towards solutions.

In this paper, we argue for a novel philosophy for biodiversity data science that embodies the spirit of the Web as "small pieces loosely joined" through the decentralized coordination of data across scientific languages and communities (Weinberger 2002). The latter, in turn, should closely approximate Wenger's (1998) *communities of practice*. Biodiversity data infrastructure and informatics tools should then reflect the forms of social engagement, imagination, and alignment within and among communities of practice that best promote biodiversity learning and high-quality data. By necessity, each community will have a taxonomic, regional, or institutional reach far below that of a single, global network. However, cross-community coordination towards a more global reach can happen simultaneously through the articulation of classification alignments enabling translation of biodiversity data across evolving or conflicting, locally maintained perspectives. Decentralized coordination therefore depends on the relatively lossless translation of data but not on a stable, shared theory of the world or pool of settled knowledge. As a result, the approach enables high-quality data synthesis without presupposing unification as a necessary starting or ending point for science, although it is consistent with both.

The contrast we draw between decentralized coordination and centralized unification has broader consequences for biodiversity data governance. Future success at modeling and managing risks to biodiversity on a planetary scale will depend on our ability to integrate diverse primary biodiversity data sources to model complex interactions between taxonomic groups and their physical environments, including humans (IPBES 2019). In this regard, we should understand and evaluate every technical solution in data science as also being a social intervention in how our scientific and public communities operate (Gerson 2008, Edwards et al. 2013). Rather than building a top-heavy global infrastructure with weak connections to existing institutions and efforts at local and regional scales, we need to innovate and promote competing designs to empower and coordinate both existing and newly emerging communities of practice. We therefore end by reviewing emerging solutions and next steps toward decentralized yet globally coordinated biodiversity data services.

2. Global biodiversity data from a knowledge commons perspective

As it has become clear that humans are fundamentally altering the face of life on the planet, many countries and international organizations have prioritized global monitoring for human impacts on biodiversity and invested in major initiatives aiming to scale up scientific knowledge and research by pooling existing resources and launching major new infrastructure projects (Adams et al. 2002). Highly centralized control may appear to be necessary for building successful shared infrastructure at such large scales, but this is not borne out empirically (Hess and Ostrom 2006, 2007, Frischmann et al. 2014, Strandburg et al. 2017). In fact, centralization can *undercut* the ability of individuals and local communities to function and contribute valuable information, expertise, and time. In this section, we highlight the important role that communities of practice play in the success of knowledge commons, especially in contexts with decentralized governance.

The idea of a "commons" refers to an institutionalized arrangement for the use and management of a shared resource among multiple actors (Strandburg et al. 2017). A commons therefore denotes "a form of community management" for a resource; not the resources, community, or place being managed (Strandburg et al. 2017, 10). While the term is also often used in a loose metaphorical way, several types of commons have been subject to extensive theoretical and empirical research, including natural resource commons and knowledge commons (Ostrom 2010, Strandburg et al. 2017, Schweik and English 2012). A knowledge commons in particular is defined as "the institutionalized community governance of the sharing and, in many cases, creation of information, science, knowledge, data, and other types of intellectual and cultural resources" (Strandburg et al. 2017, 10).

There is no simple theory that tells us how to build a sustainable, shared body of scientific knowledge valued by many different stakeholders. Contrary to conclusions suggested by early game theory models, centralizing control over a shared resource pool is neither necessary nor the most common way to prevent overexploitation, i.e., to prevent the tragedy of the commons (Wilson et al. 2013). Emphasizing *governance* for commons instead of *government* therefore highlights how authority and power over a shared resource pool are distributed beyond the legal powers of the state. Of course, the state still typically plays some role — for example, as an enforcer for contracts or intellectual property rights created by actors involved in knowledge commons — but many of the rules for access, use, and contribution are created and enforced by other actors, such as professional societies, non-profit organizations running data portals, foundations, and individual scientists. Careful empirical studies are therefore essential to understand how economic, political, and social processes influence the growth and sustainability of knowledge commons. In this regard, Sabina Leonelli has argued that "the real source of innovation in current biology is the attention paid to data handling and dissemination practices and the ways in which such practices mirror economic and political modes of interaction and decision making, rather than the emergence of big data and associated methods per se" (Leonelli 2016, 1).

Big data efforts to improve the discoverability of many small to medium-sized datasets across the world have indeed had transformative impacts on the scientific community and data ecosystem. The Global Biodiversity Information Facility (GBIF), for example, now serves 1.3 billion occurrence data points for users to search and download, aggregated from a wide range of citizen science projects, museum collections, and other organizations. Similar large national or continental aggregators exist, such as Integrated Digitized Biocollections (iDigBio; Hanken 2013) and DataONE (Michener et al. 2012), as well as for specific taxa such as the Avian Knowledge Network (Iliff et al. 2009). Aggregation at this level relies heavily on broad adherence to the Darwin Core standard for publishing occurrence data, which imposes minimal metadata requirements and categories (Wieczorek et al. 2012). Relying on this minimal consensus standard, aggregator portals can then provide centralized access points for datasets sourced from a wide range of repositories, collections, and databases.

Unfortunately, simply adding more data to a problem — e.g., by aggregating observations across many sources — does not necessarily lead to improved predictions or discoveries, despite claims that suggest theoretically informed statistical analyses and experimental designs are no longer relevant (Philippe et al. 2011, Leonelli 2014, Lazer et al. 2014, Sterner and Franz 2017). As we'll see, how data aggregation happens has ramifications far beyond which data points end up grouped together in the combined dataset. The central issue will be how centralizing control over the aggregation process and resulting product limits opportunities and incentives for continuous growth and fitness-for-use of the commons.

Where totally centralized governance would put all decisions regarding actors' access, use, and contribution rights in the hands of a single organization or group, totally decentralized governance would instead give full autonomy in each of these respects to local organizations or communities (Brown and Grant 2005). There are a range of intermediates between these polar opposites, reflecting different forms of governance arrangements and the multiple aspects of actors relationships to the knowledge commons (Ostrom 2010, Fisher and Fortmann 2010). The collaborative software development service GitHub, for example, provides an intermediate option for programmers where the updating and release of a software project can be controlled by a single individual or group. But any individual is able to initiate a new version ("fork") of the code and edit it in a decentralized fashion without any obligation to re-integrate the code back into the main project.

Despite the apparent potential for exploitation or anarchy, decentralized governance can succeed when robust local institutions exist that teach, enforce, and maintain rules (either formal or informal) about appropriate actions. The concept of a "community of practice" provides a useful way of understanding how people negotiate these rules and their application over time

(Wenger 1998, 2000). A community of practice provides the interpersonal and institutional scaffolding needed for people to be recognized, trained, and leaders in a particular skill set. The possible subject matter for a community of practice ranges broadly; including for example claims processing in insurance, macro photography, and college teaching. In biodiversity science, examples would include researchers working to advance the classification of a taxonomic group such as birds, coordinating on the conservation of a particular ecosystem, or working to improve informatics tools for sharing occurrence datasets. The relevant community is restricted, though, to a particular group of people (not necessarily in one geographic place) who interact regularly as part of learning, applying, and improving these skill sets. The community is therefore connected by the practice as a shared matter of concern, such that they are involved in a joint enterprise based on mutual interactions and a shared repertoire. It forms a group to which people can belong through their engagement and alignment in joint projects using the relevant skill set and through a shared vision for the community's aims and identity.

In order to explain and potentially influence long-term outcomes for knowledge commons, such as sustainability and growth, it is therefore essential to look beyond general open science frameworks and analyze the fine-grained rules affecting how actors engage with each other and the shared resource pool on a regular basis. As we zoom in to look at how scientific data is created, shared, and curated in practice, it is clear that governance for knowledge commons operates in multiple dimensions simultaneously and may reflect heterogeneous strategies for different aspects of the data and modes of engagement. One reason is that actors can participate in the commons in multiple ways, including adding, altering, removing, restricting, using, and exchanging access rights to the shared resources (Frischmann et al. 2014).

As an example, we can look at participation in new norms around open data, such as the FAIR principles (Wilkinson et al. 2018). From a top-down policy perspective, compliance with open data principles is an obligation for individual scientists as part of their funded research. Proactive individual adoption of open data practices also matters for the success of large cyberinfrastructure projects, such as Dryad (White et al. 2008) or DataONE. Addressing these concerns, survey studies addressing biodiversity data sharing have primarily focused on scientists' individual attitudes or behaviors (e.g. Enke et al. 2012, Schmidt et al. 2016).

From a bottom-up community perspective, however, following norms about data use and sharing are one aspect of participating in the community's broader scientific identity, culture, and social organization (Wenger 1998, Strasser 2011, Aronova et al. 2010). Research has shown these community-level variables have a strong impact on the success or failure in the related domain of collaborative open source software projects (Schweik and English 2012). Some of the most successful community curation projects for genome data have also emerged as bottom-up collaborations that use wiki technology to layer information on top of centralized sequence repositories (Lee 2017).

This suggests another reason for the importance of tracking the fine-grained governance of scientific data: the data themselves are complex objects whose properties and scientific significance often vary with context (Franz and Thau 2010; Millerand et al. 2013, Leonelli 2016, Lee 2017). In order to find and use scientific data online, metadata is key, but the language scientists use to describe the observations they have made are often highly local and dependent on tacit knowledge (Leonelli 2016, Sterner and Franz 2018). One response is to centralize control over standardized definitions and rules for metadata categories within a single organizational body. In the life sciences, for example, the development of a computer ontology for data classification in a domain often involve establishing a consortium in charge of formulating and maintaining consensus definitions for the associated terminology (Smith et al. 2007, Wieczorek et al. 2012, Millerand et al. 2013).

A decentralized strategy, by contrast, would allow multiple metadata systems to coexist when important to the functioning of relevant communities of practice. However, a good decentralized strategy should also ensure that sufficient resources exist to allow efficient and accurate translation across locally variable metadata systems. This approach relies on consensus about how to coordinate across local variation among communities of practice rather than consensus about a single global best option for all such communities (Sterner et al. In Press).

Having set out a general contrast between centralized and decentralized approaches to knowledge commons, plus the critical role of communities of practice in the long-term sustainability of growth of commons, we turn to describe key limitations of the dominant, centralized governance for global biodiversity knowledge commons and then propose a constructive alternative. As an arrangement for regulating the shared resource pool of data, the decentralized approach we propose relies on everyone sharing rules about how to make data translation possible across locally variable contexts, not on agreeing to a globally standardized dataset and metadata system.

3. Reframing the critique of globally centralized biodiversity data

Arguably the primary functionality provided by current biodiversity aggregators are centralized portals for finding and downloading occurrence datasets (Hobern et al. 2019). This is clearly a desirable feature for users exploring the data available for projects outside existing thematically focused and curated collections. However, implementing this feature drives further choices between a centralized versus decentralized strategy for handling data aggregation. In this section, we critique two such choices: centralized control over the metadata categories, especially taxonomic classifications, used to aggregate data, and centralized control over the flow of information reflecting new curation or collections work.

One major and persistent problem with globally centralized biodiversity data aggregation is that it can distort the underlying signal in heterogeneous datasets, leading to a "synthesis that nobody believes in" (Franz and Sterner 2018; see also Franz et al. 2016). This will occur, for instance, whenever occurrence records from multiple low-level sources are annotated with taxonomic name usages that are coherently applied *within* one source, but are nevertheless non-congruent *across* sources. As an example, we have shown how multiple locally and simultaneously applied meanings of taxonomic names for endangered Southeastern United States orchids in the *Cleistes*/*Cleistesiopsis* complex will produce misleading ecological and conservation inferences if aggregated under one centralized classification (see also Peterson & Navarro-Sigüenza 1999).

The distortion in signal created by centralized aggregation also has a social aspect: it generates an unbalanced allocation of power to represent and propagate conflicting data signals among aggregators and communities or individual experts advancing divergent views. To illustrate this point, we need to describe the current organization of the biodiversity data ecosystem in more detail, focusing on how data are structured flow in the aggregation process. Figures 1 and 2 provide related, schematic representations of how biodiversity data are aggregated under a centralization paradigm. It is sufficient for our purposes that this schema has 2-3 distinct aggregation layers, though in practice there are often additional layers and/or more complex data flow relationships. We also note that, while certain institutional entities, software platforms, and information communities are explicitly named (reflecting a North American emphasis due largely to the authors' expertise), this is done mainly to illustrate broader phenomena in a concrete way.

Low level. At the lower levels of the aggregation hierarchy we find individual private or institutionalized collections that may or may not be digitized in some form, but in either case are not published online in a way that enables standard-compatible, occurrence-level aggregation. We may term these "hidden data" (Fig. 1: 5A, 5B, 6).

Other collections may meet the aforementioned criteria, and can be accessed online through custom publishing solutions (Fig. 1: 3A, 3B, 4); or through more widely available, open source or commercial (fee-based) software applications that are designed to serve multi-collections institutions such as museums, academic institutions, environmental organizations, etc. (e.g., the KE Emu and Specify software applications; Fig. 1: 1A–1C, 2A–2C). Typically, these providers will *not* host occurrence records in their managed systems whose corresponding specimens or vouchers are owned by external entities. Moreover, their data packages remain globally undiscoverable according to FAIR standards (Wilkinson et al. 2016) *unless* two conditions are met: (1) the data signals are accurately and consistently mapped to a metadata standard – in our case Darwin Core Archive format (DwC–A); and (2), the packages are exposed to the web via a DwC–A interface such as the Integrated Publishing Toolkit (IPT; Robertson et al. 2014). Only data package publication with the IPT or similar interfaces allows global data discovery facilitated by an Application Programming Interface (API) and further protocol-driven aggregation.

Mid level. As herein defined, mid-level aggregators are transparently *not* designed or motivated to aggregate biodiversity across all taxonomic groups or beyond (sub-)continental boundaries. These platforms do, however, have the ability to host data from collections pertaining to multi-institutional communities of practice in the sense of Wenger (2000). Such community-driven data platforms are often referred to as *portals*. Depending on the software design and implemented sustainability model, the cardinality between software solutions and portal communities may range from one-to-one – i.e., a software program supports a one-off portal applications – or one-to-many, i.e., repeated, independent portal installations using the same software platform and hosting different data collections reflecting each communities' interests and social dynamics.

Data packages may be managed in such mid-level portals in the following two ways: (1) "live-managed," which we understand here to mean that the entity owning the physical collection of specimens or vouchers has comprehensive rights *within* the pertinent portal to create new occurrence records and annotations (referred to as within-portal rights); and (2) "snapshots," which are time-stamped *versions* of a live-managed collection exported to one or more *outside* portals where occurrence records may be annotated further – though typically by actors that are not members of the physical collection-owning entity (referred to as outside-portal rights). We note that the temporal immediacy (real-time versus delayed/filtered) of record creations and annotations is in principle separate from within- and outside-portal rights, though in the current infrastructure they are linked practically.

Unless multiple entities own specimens or vouchers within a collection, an individual collection would typically only experience 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 (even 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.

Certain mid-level aggregators may only support live-managed collections (e.g., Arctos, CITE; Fig. 1: 7, 8). However, the Symbiota software platform (Gries et al. 2014) supports live-managed as well as snapshot collections in the same portal (Fig. 1: 1B, 9, 10). This portal community-sustaining content management system furthermore has a built-in DwC–A publishing module. Both design features – i.e., the mixed live-managed/snapshot capacity and the built-in DwC–A publishing tool – are essential for exchanging occurrence records and annotations *laterally* across mid-level portals. Accordingly, SEINet, a herbarium-based portal focused on Southwestern United States 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 macrosystems ecological monitoring and forecasting. Reciprocal exchange among portals with different origins, emphases, management structures, and target communities embodies a lateral rather than hierarchical network at the middle level.

High level. High-level aggregators aim for comprehensive taxonomic and geographic coverage on national, continental, or global scales; for instance all natural history collections occurrence records of North American institutions (Fig. 1: iDigBio), or even all organismal records, globally (Fig. 1: GBIF – Global Biodiversity Information Facility 2019). They typically prefer to receive occurrence data packages and periodical updates thereof directly from the lowest-level DwC–A publishing-enabled collection instance. This means that some collections may be discoverable only at this highest level, or only at the lower and high level, yet not at the middle level when the collection has no data publishing arrangement with a mid-level portal.

By design and aspiration, then, high-level aggregators aim for aggregated occurrence data services and user communities who expect a "one-stop shop" in order to get data signals for "everything there is." On the surface, this appears as a distinct advantage over the mid- and low-level providers. However, high-level aggregation also frequently incurs the disadvantage of only serving up collection snapshots. Whereas globally comprehensive data discovery is favored at this level, new data annotations are not. Instead, high-level aggregator-derived data packages are downloaded into external web or local environments, where further data quality scrutinizing and improvement, as well as data analysis and publication of outcomes, tend to occur. Because of the hierarchical aggregation design, it is exceedingly hard to "push" these value-added data annotation back down to the live-managed collections. The vast majority of new annotation only flow one way, i.e. upwards, leading to a strongly *unidirectional* as opposed to reciprocal or cyclical quality enhancement system (Fig. 2, left region).

Implications. Pushing new annotations down from the high levels under the centralized, hierarchical aggregation paradigm is hard in a technical sense. These value-added actions tend to occur in external systems, not the software platform hosting or serving the occurrence data, and DwC–A transmission channels from source to user are often configured to function unidirectionally. It is also challenging in a social context. Most biodiversity expert communities of practice 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 that are much more accurately represented at low and middle levels of aggregation (Fig. 2, right region). Conversely, research communities that *are* interested in analyzing all organisms at the global level tend to lack the highly contextualized 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 latter point may help explain why in practice centralized aggregation of occurrence records and centralized management of data-structuring biological classifications often go hand and hand. For instance, GBIF – currently the most globally comprehensive aggregator of all – is actively promoting its "backbone" taxonomy (de Jong et al. 2015) to which all aggregated data are preferentially matched (Franz and Sterner 2018), often with considerable loss of syntactic preference and semantic context. Thereby GBIF preferentially responds to the needs of those users who are interested in large-scale signals yet who are also likely unprepared or unwilling to add value to data packages live-managed at much lower levels.

We must clarify, however, that centralizing occurrence record communities and centralizing taxonomic knowledge can be decoupled. Conversely, there is value in exploring the decentralization of each component separately, as well as the coordinated decentralization of both, as we will show below. Nonetheless, the largely unidirectional flow of new annotations under the centralized scheme is in itself undesirable, as it leads to potentially infinite, variously differing and outdated versions of "the same occurrence records" stored in unconnected systems, without adequate mechanisms to ensure annotation provenance and translation of syntactic and semantic changes.

4. A model for decentralized but globally coordinated data aggregation

In the previous section, we identified several major problems to implementing a shared resource pool for biodiversity data based on centralized governance. Conversely, a decentralized strategy, while potentially better able to accommodate local communities of practice working on and with the data, risks fragmentation of the resource pool overall. Unlike some of the most familiar natural resource commons, such as aquifers or forests, data commons cannot rely on natural processes to aggregate resources into a single, shared resource. Instead, aggregation must be engineered and re-engineered over time to ensure continued coordination of new data contributions and modifications to existing datasets. This sets up a critical challenge for decentralized approaches to data aggregation: how to engineer the capacity for competing hypotheses and distributed curation work without losing the connectivity required for global data sharing?

In this section, we present a novel model for achieving decentralized but globally coordinated data commons. We also illustrate how it addresses a key part of the coordination challenge through computer ontology alignment (Euzenat and Shvaiko 2013). As we use the term here, computer ontologies provide standardized metadata vocabularies organized into formal logical relationships that allow users to categorize and reason about data in a machine-readable format. A taxonomic classification can be expressed as a computer ontology, for example, as in the NCBI Taxonomy (Federhen 2012). The best-known example in the life sciences, though, is the Gene Ontology, which categorizes experimental results and knowledge about molecules, functions, and spatial structure in cells (Gene Ontology Consortium 2017). The Gene Ontology is one member among a much larger group of the Open Biomedical Ontologies (OBO) Foundry, which represents a highly centralized approach to ontology design and maintenance (Smith et al. 2007).

Historically, the OBO Foundry has been aligned with "ontological realism," the most high-profile theory about designing ontologies for use in science, created by Barry Smith and Werner Ceusters (Arp et al. 2015). Other approaches appeal to different theoretical foundations (Seltmann et al. 2012), but what matters for us is that all these share a common view that the design of an ontology for expressing a body of data should reflect the settled consensus view of a community. Standardization of a field's metadata categories on a single reference standard can be efficient when stable consensus actually exists, but it is neither helpful nor necessary to force adoption of a global standard.

Although it has been largely overlooked in contemporary discussions of big data, biological taxonomy has in fact developed sophisticated alternative solutions to data communication and integration based on centuries of experience of managing ever-increasing volumes and rates of heterogeneous data (Ogilvie 2003, Müller-Wille and Charmontier 2012). In particular, historical naming and classification practices in the Linnaean tradition rely on a principle of coordinative rather than substantive consensus: the design of a formal classificatory system for expressing a body of data should be grounded in a consensus standard for coordinating the use of classificatory terms, even if the meanings of those terms haven't been settled (Sterner et al. In Press). The core insight is that a community can maintain multiple reference ontologies for its data if people provide sufficient information to enable accurate and efficient data communication across each view. Historically, taxonomists have made this possible among human experts through providing nomenclatural types and taxonomic concepts as part of their definitions for new scientific names. Computers have different cognitive strengths and limitations than human experts, however, so these traditional practices must be modified and recalibrated for data-intensive science (Sterner and Franz 2017).

Our proposed model incorporates historical practices enabling coordinative consensus from biological taxonomy and extends them into the world of data-intensive science by extending recent advances in collaborative software development. In particular, the Git model for decentralized version control provides a powerful and successful foundation for realizing decentralized but globally coordinated data services (Loeliger and McCullough 2012). Perhaps best known through its implementation by GitHub, the Git model allows a group of software developers to create parallel versions ("forking") of a shared reference standard (the "master") and edit these versions locally before merging the edits with the reference standard (via a "pull request"), which may itself have changed in the meantime. Similarly, local versions can be updated with changes from the reference standard (a "push") by reconciling edits to the local and reference versions. Adopting the Git model for a project comes with implicit governance decisions about who has the ability to create local versions, request and approve changes to the reference standard, and push updates from the standard to local versions. Contributors to a collaborative project will generally form a community of practice, and the appropriate governance strategy within a community can vary from highly centralized to highly decentralized; indeed, communities often evolve over time as they grow or change identity (Shaikh and Henfridsson 2017).

The existing Git model has a couple key limitations, however, which need to be addressed for achieving decentralized but globally coordinated biodiversity data. One critical limitation of current implementations is that while they can track line edits to documents, they do not evaluate the semantic implications of those edits. In the context of software development, this means that updating the reference standard with changes in a local version does not take into account any consequences to the program's input-output behaviors. (There are heuristics to flag changes that may create problems, e.g. parallel changes across versions to nearby lines of code, but these flags are based on proximity in the document, not effects on run-time behavior.) In the context of data aggregation, this means that we could tell when the taxonomic authority for a specimen's nomenclatural identification has been changed but not necessarily whether this involves a change in the associated meaning of the taxon name.) It is therefore essential to extend the basic Git model to incorporate semantically-aware conflict detection and reconciliation between datasets with different metadata classification systems (Arndt et al. 2019).

A second limitation is that semantically-aware data reconciliation needs to be possible across multiple reference standards rather than solely with respect to local versions of a single reference standard. A setting where systematists maintain multiple, partially conflicting species checklists (e.g. as has been the case with birds for decades) is most analogous to handling decentralized versioning across multiple Git projects (i.e. multiple “master” versions), each of which has its own local versions. Aligning concepts rather than text documents (i.e. the meanings of metadata terms rather than the documents specifying them) is therefore doubly essential for accurate and machine-automated data aggregation across parallel metadata systems.

Figure 3 illustrates how our proposed model extends the Git approach to accommodate semantic alignments of multiple reference standards and thus aggregation of associated datasets. Franz et al. (2016a) have shown how such semantically-aware data translation is possible using logic reasoning based on expert-articulated relationships between metadata categories using Region Connection Calculus-5 (RCC-5). While the name sounds forbidding, RCC-5 can be understood intuitively as describing five types of relationships between the set of members associated with any pair of concepts. Given two concepts, A and B, RCC-5 lets us say whether the members of A are the same as B (==), a subset of B (<), a superset of B (>), overlapping with B (<>), or non-overlapping with B (!). Articulating the relationships between classifications using these relationships enables a computer to reason about how instances of a concept in one classification can be translated into instances of concepts in another classification. A computer is also able to check whether an alignment between two classifications is logically consistent (no contradictory statements about how to translate the data) and comprehensive (every instance of every concept in each classification can be mapped in an unambiguous way onto the other classification).

Figure 4 then shows how such alignments figure into achieving decentralized data coordination. A community of systematists working on the taxonomy of a group will typically generate multiple proposed classifications over time. Some of these classifications will prove more popular in use for data curation, shown in the figure by the width of the lines. In order to assemble a global data commons, one has to be able to access a comprehensive aggregate of existing data under a single, coherent metadata language. As we discussed with coordinative consensus, this does not entail that all data curation must be done under a single consensus classification. Instead, we need appropriate information about the meanings of the metadata categories within each classification in order to translate across them in an accurate and efficient way. While automated solutions are desirable, high-precision methods for aligning ontologies and detecting conflict remains a challenging research front in computer science. It is therefore likely that expert-curated alignments will continue to be necessary for situations where accurate data translation is important for data commons participants.

In sum, lateral alignments across metadata reference standards used by data portals and individual experts can remove the need for a single, globally authoritative consensus classification. Automated reasoning with logical relationships such as in RCC-5 lowers barriers to communities developing local consensus classifications customized to their aims and resources. Data portals primarily serving scientists outside systematics, such as conservation biologists or ecologists, may prioritize a more conservative approach to data aggregation based on the most widely used classifications. Alternatively, portals with high levels of crowd-sourced curation may find it valuable to allow contributors to use their preferred classification system out of several options (e.g. iNaturalist 2019). As Figure 4 suggests, though, alignments between some metadata languages will be more valuable for global data coordination than others. This reflects an important way in which our approach avoids having to pay the labor cost of aligning everything to everything: communities of practice or individual experts are incentivized to connect their preferred datasets and metadata languages to widely shared reference standards, but not at the cost of having their research products overridden or coarse-grained as part of participating in the knowledge commons.

5. Governance for lateral data flow among community portals

Figure 5 outlines one suggested solution to manage biodiversity data aggregation in a global, decentralized system. Again, we have chosen to present this schema with reference to concrete portals and software solutions (i.e., Symbiota; see Gries et al. 2014). However, our key interest is to make the case for a general solution path that can and should be implemented across many different software and data communities.

Our proposed solution asserts that, once there is a potential network of portals whose respective foci and boundaries correspond to active and thriving communities of practice, the task of globally integrating these portals can be achieved through newly developed and successively refined portal-to-portal Application Programming Interface (API) services. The function of the API services is precisely that of "hard-wiring" the kinds of meta-/data access-specific rules and opportunities that maintain the social integrity of each portal community while allowing filtered data annotations (Morris et al. 2013) to flow globally.

At present, lateral data flow between (Symbiota) mid-level portals, and hence also between live-managed and snapshot collections represented in these portals, is still largely facilitated through batch updates of DwC–A packages that are triggered periodically but also manually, i.e., by an individual (human) with appropriate access rights. This is precisely the fulcrum where the new decentralized commons must be realized. A technically and socially sophisticated cross-portal API data flow culture should achieve the following.

- Provide continuous, bi- and multi-directional, API-managed data flow across portals and across live-managed versus snapshot collections.
- Facilitate API-driven *filtering* of this data flow in all (likely evolving) ways needed to represent the social agendas of distributed data and data annotation providers and users. For instance, a portal focused primarily on occurrence data from Panama and frequently physically represented at the Smithsonian Tropical Research Institute (Fig. 5: STRI) we need *only* that subset of occurrence records from a portal of Southwestern United States herbaria (Fig. 5: SEINet) which represents plant occurrences from Panama. Conversely, the SEINet portal will only require accessing and ingesting data annotations by members of the STRI portal *if* they pertain to SEINet occurrences. How exactly these reciprocal updates are pushed from portal to portal, and how directly they "replace" - with provenance - the data annotations present in the live-managed collection, is again regulated directly through the API service configuration. As an example, data annotations provided on snapshot occurrences by a highly accomplished and trusted expert of a particular taxonomic group and for a particular region (filter: {individual, taxonomic group, region, period of annotations, etc.}), can be allowed through the API configuration to directly become pushed to replace standing data and annotations in corresponding the live-managed collection. In other instances, managers of the latter may have a preference to review such annotations prior to assigning them "latest viable version" in the live-managed portal. This may also entail the option to display the new annotations separately in the live-managed collection yet in a linked from, as an alternative view (so to speak) that is however not validated by the live-collection manager(s). In this sense, the API-mediated filtered on reciprocal data mirror the roles and aspirations of individual members and social entities in respective community of practice.

- Automate API-configured portal-to-portal updates such that they are triggered on a regular basis (say, each night). This design feature of a decentralized system will increasingly *do away* with a currently effective, yet also undesirable consequence of the live-managed versus snapshot distinction, i.e. that the latter (snapshots) often lag weeks or months (or more) behind the live-managed collections in terms of representing new data annotations applied. However, automatically scheduled network-wide updates would allow all new annotations to be propagated globally (though only laterally; see Fig. 5).and in accordance with (largely) one-time established rules for propagation and filtering. We note in passing that such application-to-application synchronization has become standard within the digital world. For instance, data queries, bookmarks, and messages are regularly synchronized between individuals phones, laptops, and desktop systems. In this same manner, we can synchronize occurrence datasets across a decentralized biodiversity data community network.

Based on these new boundary conditions, we can re-integrate individual researchers and teams into this decentralized network in the following ways, thereby also avoiding the externalization of their new contributions - a consequence of the hierarchical model (Fig. 2). We are provisionally developing these services under the project Symbiota/BioCache (though again, names and specific prototype implementations matter less than the deeper arguments).

The core concept of the BioCache model is to provide data discovery tools that allow research teams to harness an API-enabled network of data providers, and to define a remote index of occurrence records that will be used to address a particular research project. This data index represent a local versioning of the original, distributed data, and provides a platform where the data can be analyzed, modified, and extended - e.g., error correction, adding of coordinates, occurrence record re-/identification according to a preferred classification, annotation of misidentifications, etc. - in order to establish a locally robust dataset deemed sufficiently fit to address a specific research interest. In a sense, the BioCache data index represents a selected "fork" of individual occurrence records analogous to how code forks are defined within GitHub.

Analyses of research datasets may take months or years, with data annotations taking place throughout the distributed data network, thus establishing a need to coordinate data modifications across the network. Research teams need the ability to ensure that their remotely managed datasets are up-to-date with the distributed network of data providers. At any point in the data package assembly and validation process, mechanisms are needed to pull in additional "upstream" modifications (analogous to a Git pull), which would add new records, append coordinates, and incorporate expert determinations that regularly take place within the collection community. The distributed network needs to be able to track data modifications based on Globally Unique Identifiers (GUIDs; see Nelson et al. 2018), timestamps, and value changes so that conflicting annotations can be identified, and possibility resolved based on a set of rule-based guided decisions.

A BioCache instance acts like a new instance of a very narrowly circumscribed community of practice and, critically, adheres to the same commons of much larger and less ephemeral communities such as portal. Data annotations made in the BioCache instance are thus accessible to the original data source via the API-driven infrastructure built into the instance's data indexer. This design feature provides the means for the original data provider to selectively pull annotations back into the original, live-managed occurrence record, or at least to define a reference to the extended data product. However, each occurrence record therefore represents its own version fragment that will need to be evaluated, resolved, and merged back into the original data repository on its own terms. In contrast to a Git code fork, a typical BioCache dataset will consist of occurrences harvested from numerous live-managed portals. Negotiations for a "data fork" to be merged back into the source portals should involve a variety of evaluation and resolution criteria uniquely defined by the various agents owning the data. Conversely, research teams should have the means to submit "pull requests" from the remote data fork to all live-managing data providers at once. The latter are offered several options for integrating data annotations back into their live-managed collections and portal. Ideally, occurrence record management tools are enabled with evaluation and resolution mechanisms that can access the research instance's API services to selectively ingest the annotations on the specific terms of the live-managed collection.

We regard this model as a critical precondition for establishing and nurturing local, high-quality data-preserving communities of practice for both occurrence records and taxonomic knowledge systems. Many preconditions for such a decentralized network are already in place. Current technical limitations are largely rooted in a certain level of isolation that exists across portals. A robust infrastructure to support a broad range of data synchronization function across portals is not yet realized. Further tracking of data provenance, versioning, and the establishment of API-facilitated data transfer protocols can provide an enhanced level of social relations across portals. This infrastructure would allow expert knowledge managed within one portal to be shared and coordinated across the portal network. Domain knowledge expertise would freely flow out into the extended specimen network with external annotations channeled back to the domain expert for review, verification, and acceptance. For instance, expert curated occurrence records managed within the SEINet portal can be republished as a synchronized data package within the STRI portal (Fig. 5). The decentralized model requires establishing an extended, potentially global network of API-enabled data providers, including individual curators.

6. Conclusion

As first-generation projects have matured, the emerging global biodiversity knowledge commons is reaching an important turning point. Globalizing access to occurrence data has generated major successes for science and decision-making, including improved discoverability through aggregator portals and growing adoption of minimum metadata standards such as Darwin Core. However, important limitations have become clear to a top-down strategy emphasizing centralized control over data aggregation and metadata systems. National and international biodiversity aggregators rely on a robust ecosystem of smaller communities of practice focused on particular research themes or collections, but the needs, views, and aims of these communities are at best imperfectly reflected at the global scale. The project of globalizing biodiversity knowledge must minimize harm to these communities' ability to experiment, specialize, and pursue the goals of local stakeholders. We've shown how innovative solutions to biodiversity data sharing and aggregation can underwrite greater decentralization without sacrificing efficient coordination at the global scale. We've also provided a framework for how our approach can generalize to address the need for stronger integration between decision-making, modeling, and biodiversity data sources.

Looking forward, decentralized coordination has broader potential to strengthen integration across the biodiversity data life cycle. Reflecting the societal importance and urgency of continued biodiversity loss, we characterize the biodiversity data life cycle here in terms of a positive feedback loop connecting primary data sources and decision-making. Figure 6 illustrates key components and interactions we identify as central to integrating the biodiversity data life cycle from primary data collection to decision-making. The figure can be interpreted as a schematic diagram to be filled in with particulars for a given situation, e.g. with the design of a new protected area as the decision to be made, models for species richness and spatial niches as the predictive models, and GIS data layers as the modeling knowledge base. It can also apply to many other research areas, including risk management for invasive species, emerging diseases, or ecosystem tipping points. The life cycle diagram therefore illustrates the goal of achieving a positive feedback loop coordinating decision-makers' needs with scientists' research efforts and incentivizing community data integration and modeling to yield high-quality and relevant data signals. The future of biodiversity rests, in this regard, on our ability to find new ways for computers to help scientists and decision-makers resolve complex collective action problems, and we should keep in mind that every technological fix is also a social intervention.

Figure 6 serves to highlight a further key difference in orientation between decentralized coordination and the existing one-way data flow toward high-level aggregators. Our model emphasizes the development of local communities of practice and data infrastructure aimed at impacting concrete decision-making contexts, rather than a highly standardized, global data product intended to be fit-for-use across most contexts. Scientists are increasingly recognizing the limitations of this latter, loading-dock model for scientific knowledge as a way to influence action (e.g. Cash et al. 2006): simply making all the data available for use in one place is not sufficient to bridge the gap to practical impact, especially if centralization undermines key dimensions of fitness-for-use for decision makers. Similarly, research analyzing how data travel effectively across local contexts consistently demonstrate the substantial effort required to design fit-for-use data packaging and reinterpret external datasets for local projects (Gerson 2008, Leonelli 2016). The costs of centralized data aggregation also undermine its long-term value for users, including government stakeholders and conservation planners. Signal distortion due to misaggregation across conflicting taxonomic concepts, for example, has high-stakes implications for species-level assessments of extinction risks.

Peter Galison's concept of "trading zones" may provide a helpful starting point for conceptualizing how the interactions between layers are organized and can be strengthened through improved communication and infrastructure (Gorman 2010). Trading zones arise as sites of interaction between communities with heterogeneous languages and cultures and whose members typically exchange objects or information without fully understanding or accepting the views of their trading partners. Sustained and efficient cooperation is then possible without unification under a consensus worldview or goal, although in some cases trading zones do transform into coherent research fields in their own right (biophysics is a commonly cited example). However, the heterogeneity of participants and decentralized governance of trading zones do create some persistent challenges for creating and maintaining value for parties on all sides of the exchange.

Importantly, there are two challenges to building data infrastructure for a network of trading zones. The first-order problem is making the actual exchange of data easier, but the second-order problem is facilitating how people arrange to exchange data. We've argued for the practicality and desirability of providing local data editing rights to communities of practice while maintaining the ability to distribute these updates in a coordinated fashion (i.e. to "trade" in curated datasets; see also Lee 2017). In a centralized paradigm that forces one-way data flow from sources to users via aggregators, important inefficiencies arise when customization work done to situate data for local use cannot be shared with others. We've also argued for the importance of facilitating how communities of practice transition from informal social networks to formalized governance and infrastructure. A major contributor to the success of Symbiota has been the social component of the application design. Individuals can establish research datasets within their own password-protected user space while maintaining full control over data attribution and management, however, their data is also tightly integrated into a social system which aligns with their research identity. Maintaining full control and ownership of data combined with the ability to author expert curated content to a community of knowledge experts has been a strong motivating factor to publish data within the Symbiota portal network. Taking into account both first and second-order challenges for governance seems crucial for achieving a sustainable and maximally valuable biodiversity knowledge commons at the global scale.

Acknowledgements

We would like to acknowledge several funding sources that have supported this work along with discussions with David Remsen, Anne Thessen, Katja Seltmann, and Atriya Sen.

References

Adams, Jonathan, Frank Biasi, Colin Bibby, and Martin Sneary. 2002. "The Biodiversity Knowledge Commons." *Conservation in Practice* 3 (4): 41–44.

Alphandéry, Pierre, and Agnès Fortier. 2010. "Local Settings and Biodiversity." *Current Sociology* 58 (5): 755–76.

Aronova, Elena, Karen S Baker, and Naomi Oreskes. 2010. "Big Science and Big Data in Biology" *Hist Stud Nat Sci* 40 (2): 183–224.

Arndt, Natanael, Patrick Naumann, Norman Radtke, et al. 2019. "Decentralized Collaborative Knowledge Management Using Git." *Journal of Web Semantics* 54: 29–47.

Arp, Robert, Barry Smith, and Andrew D Spear. 2015. *Building Ontologies with Basic Formal Ontology*. Cambridge, MA: MIT Press.

Brown, Allen E, and Gerald G Grant. 2005. "Framing the Frameworks: a Review of IT Governance Research." *Communications of the Association for Information Systems* 15: 696–712.

Cash, David W, Jonathan C Borck, and Anthony G Patt. 2006. "Countering the Loading-Dock Approach to Linking Science and Decision Making." *Science, Technology & Human Values* 31 (4): 465–94.

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

Devictor, Vincent, and Bernadette Bensaude-Vincent. 2016. "From Ecological Records to Big Data." *History and Philosophy of the Life Sciences* 38 (4): 13.

Enke, Neela, Anne Thessen, Kerstin Bach, et al. 2012. "The User's View on Biodiversity Data Sharing." *Ecological Informatics* 11: 25–33.

Edwards, P N, S J Jackson, M K Chalmers, et al. 2013. *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges*. Ann Arbor, MI: Deep Blue.

Euzenat, Jérôme, and Pavel Shvaiko. 2013. *Ontology Matching*. 2nd ed. Berlin: Springer-Verlag.

Federhen, Scott. 2012. "The NCBI Taxonomy Database." *Nucleic Acids Research* 40: D136–43.

Fisher, Joshua B, and Louise Fortmann. 2010. "Governing the Data Commons: Policy, Practice, and the Advancement of Science." *Information & Management* 47 (4): 237–45.

Franz, Nico M, and David Thau. 2010. "Biological Taxonomy and Ontology Development." *Biodiversity Informatics* 7 (1): 45–66.

Franz, Nico M, Naomi M Pier, Deeann M Reeder, et al. 2016a. "Two Influential Primate Classifications Logically Aligned." *Systematic Biology* 65 (4): 561–82.

Franz, Nico M, Edward Gilbert, Bertram Ludäscher, and Alan Weakley. 2016b. "Controlling the Taxonomic Variable." *Research Ideas and Outcomes* 2: e10610.

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

Frischmann, Brett M, Katherine J Strandburg, and Michael J Madison, eds. 2014. *Governing Knowledge Commons*. Oxford: Oxford University Press.

Gene Ontology Consortium. 2017. "Expansion of the Gene Ontology Knowledgebase and Resources." *Nucleic Acids Research* 45 (D1): D331–38.

Global Biodiversity Information Facility. 2019. "What Is GBIF?" Accessed November 3. URL: https://www.gbif.org/what-is-gbif.

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

Gerson, Elihu M. 2008. “Reach, Bracket, and the Limits of Rationalized Coordination.” In *Resources, Co-Evolution and Artifacts*, edited by Mark S Ackerman, Christine A Halverson, Thomas Erickson, and Wendy A Kellogg, 193–220. London: Springer.

Gorman, Michael E, ed. 2010. *Trading Zones and Interactional Expertise*. Cambridge, MA: MIT Press.

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

Hanken, James. 2013. “Biodiversity Online: Toward a Network Integrated Biocollections Alliance.” *BioScience* 63 (10): 789–90.

Hinchliff, Cody E, Stephen A Smith, et al. 2015. “Synthesis of Phylogeny and Taxonomy Into a Comprehensive Tree of Life.” *PNAS* 112 (41): 12764–69.

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

Hortal, Joaquín, Francesco de Bello, José Alexandre F Diniz-Filho, et al. 2015. “Seven Shortfalls That Beset Large-Scale Knowledge of Biodiversity.” *Annual Review of Ecology, Evolution, and Systematics* 46 (1): 523–49.

Hess, Charlotte, and Elinor Ostrom. 2006. “A Framework for Analysing the Microbiological Commons.” *International Social Science Journal* 58 (188): 335–49.

Hess, Charlotte, and Elinor Ostrom, eds. 2007. *Understanding Knowledge as a Commons*. Cambridge, MA: MIT Press.

iNaturalist. 2019. “Taxon Frameworks · iNaturalist.org.” *Inaturalist.org*. March 29. https://www.inaturalist.org/pages/taxon_frameworks. Accessed Nov 3, 2019.

Iliff, Marshall, Leo Salas, Ernesto Ruelas Inzunza, et al. 2009. “The Avian Knowledge Network.” In: Rich, T.D., C. Arizmendi, D. Demarest and C. Thompson [Eds.]. 2009. *Tundra to Tropics: Connecting Birds, Habitats and People,* 365–373.

IPBES. 2019. Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. IPBES secretariat, Bonn, Germany.

Kelling, Steve, Wesley M Hochachka, Daniel Fink, et al. 2009. “Data-Intensive Science: A New Paradigm for Biodiversity Studies.” *BioScience* 59 (7): 613–20.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. “The Parable of Google Flu: Traps in Big Data Analysis.” *Science* 343 (6176): 1203–5.

Lee, Peter. 2017. “Centralization, Fragmentation, and Replication in the Genomic Data Commons.” In *Governing Medical Knowledge Commons*, edited by Katherine J Strandburg, Brett M Frischmann, and Michael J Madison. Cambridge: Cambridge University Press.

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

Loeliger, J, and M McCullough. 2012. *Version Control with Git*. Sebastopol, CA: O'Reilly Media.

Mastrángelo, Matías E, Natalia Pérez-Harguindeguy, Lucas Enrico, et al. 2019. "Key Knowledge Gaps to Achieve Global Sustainability Goals." *Nature Sustainability* 113: 1–7.

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.

Morris, Robert A, Lei Dou, James Hanken, et al. 2013. "Semantic Annotation of Mutable Data." *PLoS ONE* 8 (11): e76093.

Müller-Wille, Staffan, and Isabelle Charmantier. 2012. "Natural History and Information Overload." *Studies in the History and Philosophy of Biological and Biomedical Sciences* 43 (1): 4–15.

Nelson, Gil, Patrick Sweeney, and Edward Gilbert. 2018. "Use of Globally Unique Identifiers (GUIDs) to Link Herbarium Specimen Records to Physical Specimens." *Applications in Plant Sciences* 6 (2): e1027.

Ogilvie, Brian W. 2003. "The Many Books of Nature." *Journal of the History of Ideas* 64 (1): 29–40.

Ostrom, Elinor. 2010. "Beyond Markets and States: Polycentric Governance of Complex Economic Systems." *American Economic Review* 100 (3): 641–72.

Peterson, A Townsend, and Adolfo G Navarro Sigüenza. 1999. "Alternate Species Concepts as Bases for Determining Priority Conservation Areas." *Conservation Biology* 13 (2): 427–31.

Peterson, A Townsend, Sandra Knapp, Robert Guralnick, et al. 2010. "The Big Questions for Biodiversity Informatics." *Systematics and Biodiversity* 8 (2): 159–68.

Peterson, A Townsend, and Jorge Soberón. 2017. "Essential Biodiversity Variables Are Not Global." *Biodiversity and Conservation* 27 (5): 1277–88.

Philippe, Hervé, Henner Brinkmann, Dennis V Lavrov, et al. 2011. "Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough." *PLoS Biology* 9 (3): e1000602.

Rittel, Horst W J, and Melvin M Webber. 1973. "Dilemmas in a General Theory of Planning." *Policy Sciences* 4 (2): 155–69.

Robertson, Tim, Markus Döring, Robert Guralnick, et al. 2014. "The GBIF Integrated Publishing Toolkit." *PLoS ONE* 9 (8): e102623.

Ruggiero, Michael A, Dennis P Gordon, Thomas M Orrell, et al. 2015. "A Higher Level Classification of All Living Organisms." *PLoS ONE* 10 (4): e0119248.

Schmidt, Birgit, Birgit Gemeinholzer, and Andrew Treloar. 2016. "Open Data in Global Environmental Research." *PLoS ONE*: e0146695.

Schweik, Charles M, and Robert C English. 2012. *Internet Success*. Cambridge, MA: MIT Press.

Seltmann, K, A Austin, and J Jennings. 2012. "A Hymenopterists' Guide to the Hymenoptera Anatomy Ontology." *Journal of Hymenoptera Research* 27 (2): 67–88.

Shaikh, Maha, and Ola Henfridsson. 2017. “Governing Open Source Software Through Coordination Processes.” *Information and Organization* 27 (2): 116–35.

Smith, Barry, Michael Ashburner, Cornelius Rosse, et al. 2007. “The OBO Foundry.” *Nature Biotechnology* 25 (11): 1251–55.

Sterner, Beckett W, and Nico M Franz. 2017. “Taxonomy for Humans or Computers?” *Biological Theory* 12 (2): 99–111.

Sterner, Beckett W, Joeri Witteveen, and Nico M Franz. In Press. “Coordination Instead of Consensus Classifications.” *History and Philosophy of the Life Sciences*.

Strandburg, Katherine J, Brett M Frischmann, and Michael J Madison, eds. 2017. *Governing Medical Knowledge Commons*. Cambridge: Cambridge University Press.

Strasser, Bruno J. 2011. “The Experimenter’s Museum.” *Isis* 102 (1): 60–96.

Turnhout, Esther, and Susan Boonman-Berson. 2011. “Databases, Scaling Practices, and the Globalization of Biodiversity.” *Ecology and Society* 16 (1).

Turnhout, Esther, Katja Neves, and Elisa de Lijster. 2014. “‘Measurementality’ in Biodiversity Governance.” *Environment and Planning A* 46 (3): 581–97.

Weinberger, David. 2008. *Small Pieces Loosely Joined*. New York: Basic Books.

Wenger, Etienne. 1998. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge: Cambridge University Press.

Wenger, Etienne. 2000. “Communities of Practice and Social Learning Systems.” *Organization* 7 (2): 225–46.

White, Hollie C, Sarah Carrier, Abbey Thompson, Jane Greenberg, and Ryan Scherle. 2008. “The Dryad Data Repository.” In *Proc Int'l Conf. on Dublin Core and Metadata Applications 2008*, 157–62.

Whitman, William B. 2015. “Genome Sequences as the Type Material for Taxonomic Descriptions of Prokaryotes.” *Systematic and Applied Microbiology* 38 (4): 217–22.

Wieczorek, John, David Bloom, Robert Guralnick, et al. 2012. “Darwin Core: An Evolving Community-Developed Biodiversity Data Standard.” *PLoS ONE* 7 (1): e29715.

Wilkinson, Mark D, Michel Dumontier, Ijsbrand Jan Aalbersberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (1): e1002295.

Wilson, David Sloan, Elinor Ostrom, and Michael E Cox. 2013. “Generalizing the Core Design Principles for the Efficacy of Groups.” *Journal of Economic Behavior & Organization* 90: S21–S32.

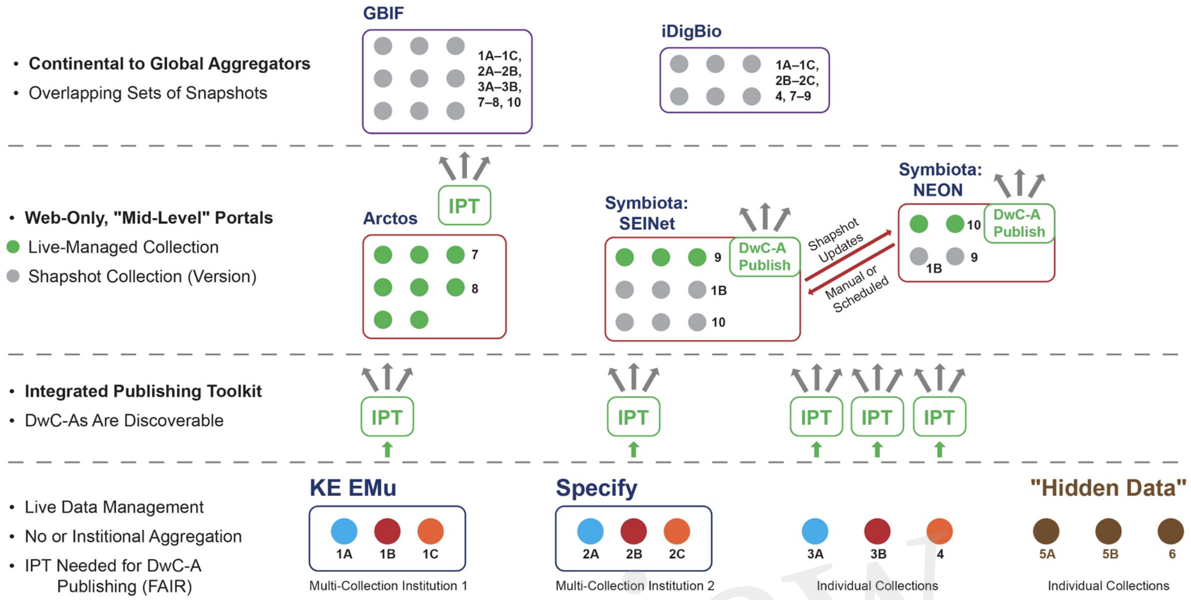
GBIF
1A–1C, 2A–2B, 3A–3B, 7–8, 10
iDigBio
1A–1C, 2B–2C, 4, 7–9
• Continental to Global Aggregators
• Overlapping Sets of Snapshots
• Web-Only, "Mid-Level" Portals
● Live-Managed Collection
● Shapshot Collection (Version)
Arctos
IPT
7
8
Symbiota: SEINet
9
1B
10
DwC-A Publish
Shapshot Updates
Manual or Scheduled
Symbiota: NEON
10
9
1B
DwC-A Publish
• Integrated Publishing Toolkit
• DwC-As Are Discoverable
IPT
IPT
IPT
IPT
IPT
• Live Data Management
• No or Institional Aggregation
• IPT Needed for DwC-A Publishing (FAIR)
KE EMu
1A
1B
1C
Multi-Collection Institution 1
Specify
2A
2B
2C
Multi-Collection Institution 2
3A
3B
4
Individual Collections
"Hidden Data"
5A
5B
6
Individual Collections

Figure 1. Schematic representation of biodiversity data aggregation under a centralized system. Collections pertaining to the same academic institution carry the same number with succeeding letters (1A, 1B, etc.). See text for additional explanation.

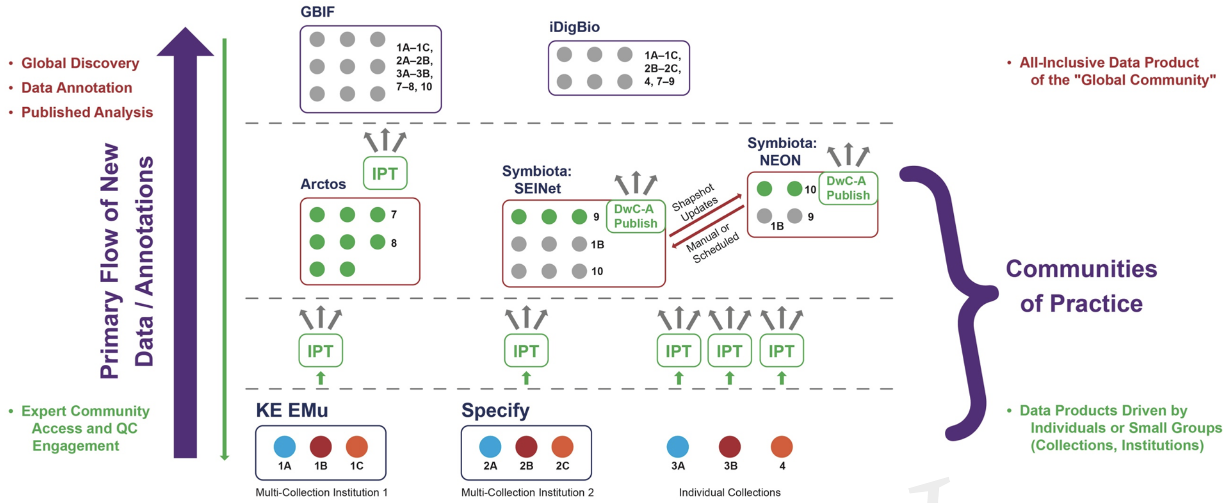
GBIF
1A–1C, 2A–2B, 3A–3B, 7–8, 10
iDigBio
1A–1C, 2B–2C, 4, 7–9
• Global Discovery
• Data Annotation
• Published Analysis
Primary Flow of New Data / Annotations
• All-Inclusive Data Product of the "Global Community"
IPT
Arctos
7
8
Symbiota: SEINet
9
1B
10
DwC-A Publish
Shapshot Updates
Manual or Scheduled
Symbiota: NEON
10
1B
9
DwC-A Publish
Communities of Practice
IPT
IPT
IPT
IPT
IPT
• Expert Community Access and QC Engagement
KE EMu
1A
1B
1C
Multi-Collection Institution 1
Specify
2A
2B
2C
Multi-Collection Institution 2
3A
3B
4
Individual Collections
• Data Products Driven by Individuals or Small Groups (Collections, Institutions)

Figure 2. Primary flow or directionality of new data items (occurrence records) and annotations under the centralized aggregation paradigm. Whereas much of the global data discovery and further analysis requires interaction at the highest level (left region), engagement with communities of practice is largely constrained to the low- and mid-level environments. See text for additional explanation.

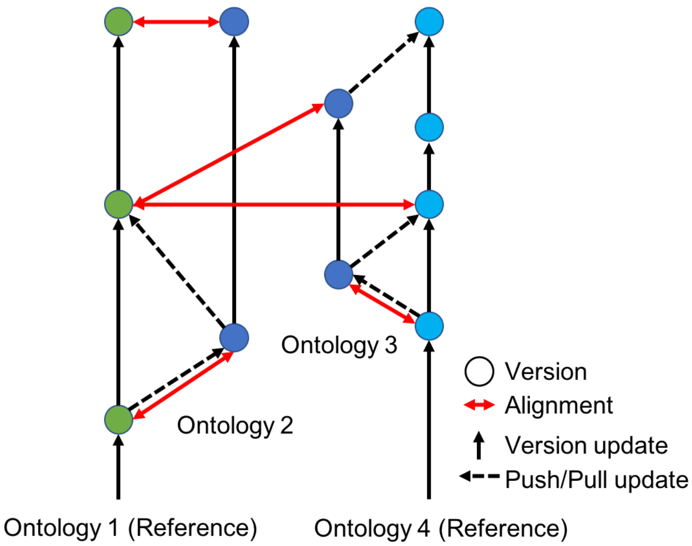

Figure 3. An illustration of semantically-aware decentralized version control for computer ontologies. Ontologies 1 and 2 serve as reference standards (green and light blue dots) shared by different communities of practice, e.g. two taxonomic classifications developed by systematists in North American and Asia for the same perceived group of species. Individual experts or subcommunities can then create local versions of these reference standards (dark blue dots) for research use or to meet local specifications of a government agency or conservation stakeholder. The ability to translate data accurately across local versions and reference standards is provided by ontology alignments (red arrows).

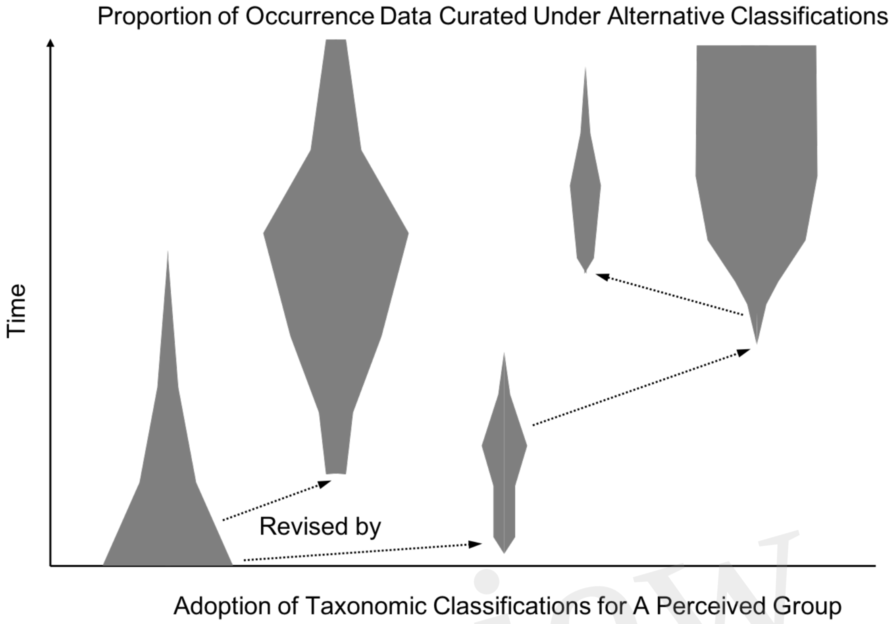
Proportion of Occurrence Data Curated Under Alternative Classifications
Time
Revised by
Adoption of Taxonomic Classifications for A Perceived Group

Figure 4. A hypothetical spindle diagram showing the relative usage of taxonomic classifications over time. Each spindle (gray polygons) corresponds to a distinct taxonomic classification for the corresponding (perceived) group, and the spindle's width indicates the relative proportion of occurrence data curated under that classification. The diagram illustrates how multiple classifications may be in heavy use simultaneously, while other proposed revisions do not become widely used but can influence future work (arrows).

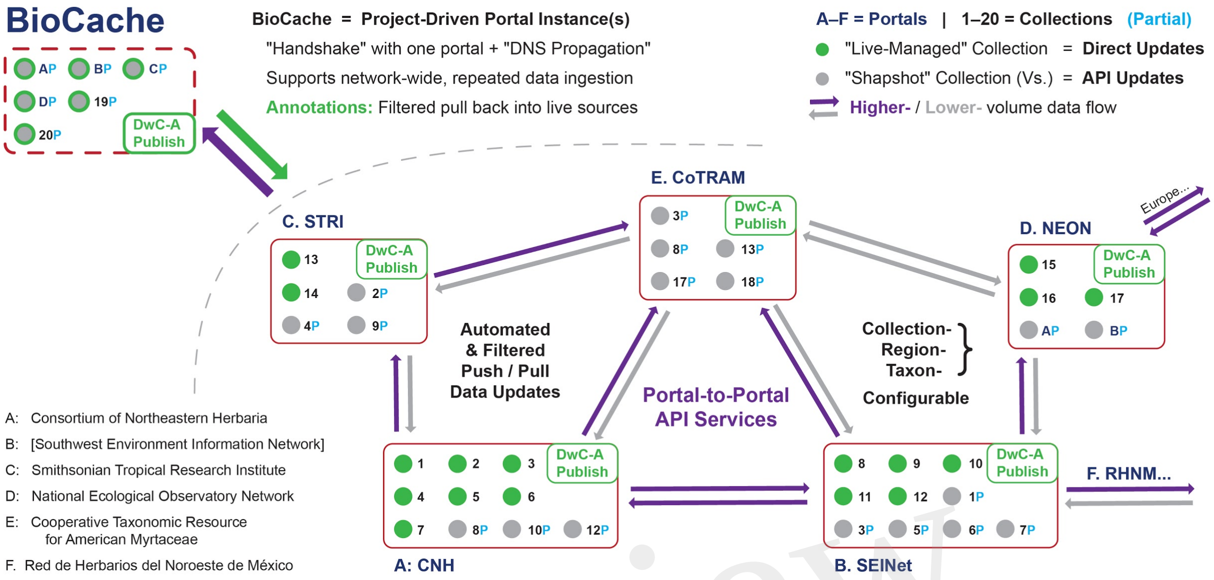
BioCache
AP BP CP
DP 19P
20P
DwC-A Publish
BioCache = Project-Driven Portal Instance(s)
"Handshake" with one portal + "DNS Propagation"
Supports network-wide, repeated data ingestion
Annotations: Filtered pull back into live sources
A–F = Portals | 1–20 = Collections (Partial)
"Live-Managed" Collection = Direct Updates
"Shapshot" Collection (Vs.) = API Updates
Higher- / Lower- volume data flow
E. CoTRAM
3P 8P 13P 17P 18P
DwC-A Publish
C. STRI
13 14 2P 4P 9P
DwC-A Publish
D. NEON
Europe...
15 16 17 AP BP
DwC-A Publish
Automated & Filtered Push / Pull Data Updates
Portal-to-Portal API Services
Collection- Region- Taxon- Configurable
A: Consortium of Northeastern Herbaria
B: [Southwest Environment Information Network]
C: Smithsonian Tropical Research Institute
D: National Ecological Observatory Network
E: Cooperative Taxonomic Resource for American Myrtaceae
F. Red de Herbarios del Noroeste de México
1 2 3 4 5 6 7 8P 10P 12P
DwC-A Publish
A: CNH
8 9 10 11 12 1P 3P 5P 6P 7P
DwC-A Publish
B. SEINet
F. RHNM...

Figure 5. Proposed solution for a decentralized biodiversity data aggregation paradigm, focusing primarily on the interaction between mid-level portal. Low-level entities as in Figure 1; however, unlike in Figure 1, the high-level aggregators are no longer relevant. As an essential feature of this proposed design, registration of the BioCache research instance is limited to one network portal; however, propagation of this registration events across the entire portal network is automatically facilitated through mechanisms analogous to Domain Name Server (DNS) propagation. See text for additional explanation.

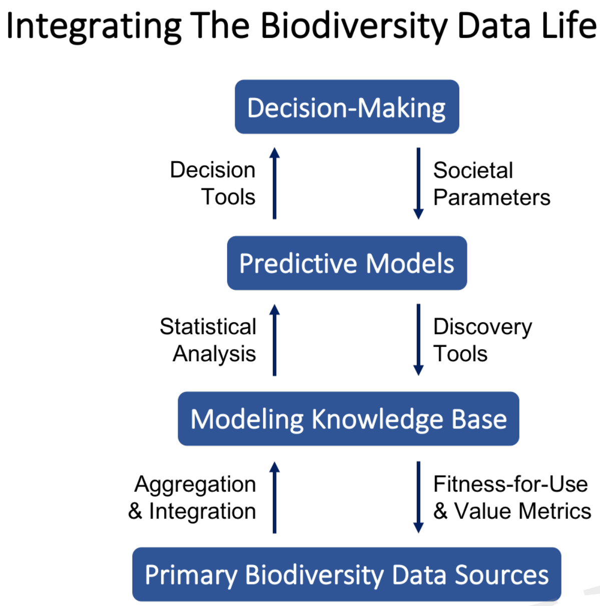
Integrating The Biodiversity Data Life Cycle
Decision-Making
Decision Tools
Societal Parameters
Predictive Models
Statistical Analysis
Discovery Tools
Modeling Knowledge Base
Aggregation & Integration
Fitness-for-Use & Value Metrics
Primary Biodiversity Data Sources

Figure 6: Each level (blue rectangles) represents a broad class of products exchanged in trading zones among researchers, decision-makers, and society members. The arrows linking levels describe information processes linking trading zones across levels.