bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

June 28, 2017

Manuscript, *bioRxiv*

Point of View

To Increase Trust, Change the Social Design Behind Aggregated Biodiversity Data

NICO M. FRANZ AND BECKETT W. STERNER

School of Life Sciences, Arizona State University, PO Box 874501, Tempe, AZ 85287-4501, USA; E-mail: nico.franz@asu.edu

Abstract.—Growing concerns about the quality of aggregated biodiversity data are lowering trust in large-scale data networks. Aggregators frequently respond to quality concerns by recommending that biologists work with original data providers to correct errors "at the source". We show that this strategy falls systematically short of a full diagnosis of the underlying causes of distrust. In particular, trust in an aggregator is not just a feature of the data signal quality provided by the aggregator, but also a consequence of the social design of the aggregation process and the resulting power balance between data contributors and aggregators. The latter have created an accountability gap by downplaying the authorship and significance of the taxonomic hierarchies – frequently called "backbones" – they generate, and which are in effect novel classification theories that operate at the core of data-structuring process. The Darwin Core standard for sharing occurrence records plays an underappreciated role in maintaining the accountability gap, because this standard lacks the syntactic structure needed to preserve the taxonomic coherence of data packages submitted for aggregation, leading to inferences that no individual source would support. Since high-quality data packages can mirror competing and conflicting classifications, i.e., unsettled systematic research, this plurality must be accommodated in the design of biodiversity data integration. Looking forward, a key directive is to develop new technical pathways and social incentives for experts to contribute directly to the validation of taxonomically coherent data packages as part of a greater, trustworthy aggregation process.

bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

Many fundamental problems in biology rely on data about the locations and traits of naturally occurring organisms classified according to taxonomic categories. These occurrence records (Wieczorek et al. 2012) are often tied to vouchered observations or specimen depositions with provenance from natural history collections, legacy publications, data repositories, and ongoing inventories. Numerous projects are bringing occurrence records into the realm of modern data science (e.g. Bisby 2000; Baker 2011; Blagoderov et al. 2012; Meyer et al. 2015; Peterson et al. 2015). This process involves multiple levels of data creation and aggregation to enable the synthesis of biodiversity data patterns and trends at greater scales (e.g. Jetz et al. 2012; Wheeler et al. 2012; Franklin et al. 2017). However, there is a widespread sense that aggregated biodiversity data should be used with caution because they are frequently of insufficient quality to support reliable inferences.

Dozens of studies have found quality shortcomings in aggregated occurrence datasets; including Soberón et al. 2002; Graham et al. 2004; Por 2007; Yesson et al. 2007; Bortolus 2008; Page 2008; Hill et al. 2010; Costello et al. 2012; Ballesteros-Mejia et al. 2013; Belbin et al. 2013; Faith et al. 2013; Gaji et al. 2013; Mesibov 2013; Otegui et al. 2013a, 2013b; Beck et al. 2014; Ferro and Flick 2015; García-Roselló et al. 2015; Harding et al. 2015; Maldonado et al. 2015; Franz et al. 2016a, 2016b; Gueta et al. 2016; Sikes et al. 2016; Smith et al. 2016; Stropp et al. 2016; Wiser 2016; Zermoglio et al. 2016; and Franklin et al. 2017. The growing concerns are leading to reduced trust in the aggregators' data, and thereby also reduce the data's scientific use and societal impact (Turel and Gefen 2015).

Biodiversity data aggregators frequently respond to quality concerns by recommending that biologists work with the original providers of the data to correct errors *at the source* (e.g., Belbin et al. 2013). This argument is exemplified by Sikes et al.'s (2016: 149) response to Ferro and Flick (2015): "[W]e have heard some taxonomists state they do not want to share data with GIBF [the Global Biodiversity Information Facility] because they distrust the quality of the data in GBIF. This latter point seems illogical. The data in GBIF are the data from the museums that provide data. If the data in GBIF are not to be trusted then neither are the data in the source museums. It thus seems illogical to be pro-museum and anti-GBIF."

We show that this strategy of allocating responsibilities for data quality falls systematically short of a full diagnosis of the underlying causes, because it fails to make sense of an important way in which many biologists have come to mistrust aggregators. In contrast, we will argue that the distrust held for biodiversity data aggregators is justified once we recognize that these aggregators have allocated responsibility in a way that holds no one accountable for the accuracy of the taxonomic hierarchies operating at the core of their data-structuring process.

On the surface, aggregators appear to recognize the need for such accountability (Belbin et al. 2013: 73): "What ideally is needed is an environment created by agencies such as GBIF and ALA [the Atlas of Living Australia] that *efficiently* enables" the exposure, discussion, and correction of "errors directly in *all relevant locations*". The authors acknowledge that "[n]o such environment currently exists", and "[p]rogress will be limited while the underlying culture of data publishing and data management does not support stable, long-term reference to each data record and community-based curation of those data in a way that ensures that each act of correcting *any aspect of any data element* is not lost but contributes to the development of a global digital biodiversity knowledgebase" (Belbin et al. 2013: 73–74; emphasis added).

Here we demonstrate that in practice aggregators have made fundamental design choices that lower trust in the aggregation process and outcome, by excluding taxonomic classification as an aspect of biodiversity knowledge worthy of being represented and preserved. Looking ahead, we

bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

need new solutions for biodiversity data aggregation that are capable of handling common situations in which research communities and information providers endorse incongruent systematic hypotheses. Based on our diagnosis, we recommend shifting the balance of power to license trustworthy data packages away from biodiversity aggregators and towards individual expert authors.

No Simple Diagnosis for Data Quality Deficiencies

It is not uncommon for researchers to think of the challenge of aggregating data within a single information system as a primarily technical endeavor without deeper scientific significance. Careful studies have shown, however, that database systems constrain as well as enable science. They have the power to channel research; e.g. by making some questions harder to ask, by entrenching certain presuppositions into their design, and by placing new burdens on participation on the research community (Bowker 2000; Strasser 2010; Edwards et al. 2011; Leonelli 2013, 2016; Millerand et al. 2013). Implementing data aggregation systems therefore places one in a position of responsibility and trust for the larger community – a status that is at times novel for researchers who by training are more focused on the domain-specific aspects of science than social engineering. Abstract notions of responsibility become concrete, however, when they intersect with the ways that an aggregation system handles the process of correcting or avoiding data quality deficiencies.

We take the presence of a growing corpus of critical studies, such as those cited above, as evidence that researchers are becoming increasingly cautious about relying on aggregators. Other biologists are bound to notice and may be deterred by the costs or expertise required to carry out the necessary additional quality vetting. That said, the helpful term "fitness-for-use" (Hill et al. 2010) reminds us that certain deficiencies can be acceptable to researchers whose inference needs are robust to them. Many biodiversity data packages and analytical needs will match up well in this sense. Nonetheless, providing data that are just clean enough for coarse-grained inferences is not a way to deal with the root issues limiting trust.

All parties have an interest in precise diagnoses of where the responsibilities for specific data deficiencies lie. Aggregators are not helped in the long run by a narrative of accountability that is too broadly formulated to allow each party to play an appropriate role in overcoming the issues. Accordingly, and given the complexity of the aggregation process, the question of who is responsible for creating and fixing deficiencies cannot have a simple, univocal answer. A common aggregation path for an occurrence record includes individuals recording the original data and metadata at the time of collection, transcribing the record from field notes into the source collection's local database, and applying further data and transformations – such as georeferencing and taxonomic identification – to comply with locally accepted conventions. Often, the collection will transfer the record to a regional, mid-level aggregator, which may then transmit the information again to a higher-level aggregrator. Alterations of 'the record' can occur at any stage along this provenance chain, and have the potential to affect the identity of the record and its empirical signal.

Responsibilities for generating and repairing quality issues are not uniformly distributed across the data aggregation chain. Correcting a false collecting date would typically be the responsibility of individuals with knowledge of the original collecting event. Dealing with repeated duplications of records through aggregation, in turn, is a task better suited for aggregators. Two examples illustrate this point. First, in their response to Mesibov's (2013) audit

bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

of aggregated data on Australian millipedes, Belbin et al. (2013) – representing the aggregator's perspective – provide a table with 44 automated quality checks performed by ALA (Belbin and Williams, 2016). The five most frequent categories of error, respectively affecting 18.6–90.2% of the records, are all related to georeferencing. Second, Hjarding et al. (2015), in their assessment of records of East African chameleons served by GBIF (Edwards 2004), conclude that 99.9% "used outdated taxonomy" that would have led to inadequate threat category assignments for eight taxa.

To extend the point about diagnostic precision further, Darwin Core (DwC), the prevailing global standard for sharing occurrence records, has seven main categories (Wieczorek et al. 2012). Of these, Event (when), Location (where), and Taxon (what) are the primary data blocks where insufficient quality will affect inferences of biodiversity distributions. How abundant and significant the shortcomings are varies from case to case.

Importance of Trust for Sustained Use

Trust is a complex and context-sensitive concept (Hardwig 1991; Fricker 2007; Sperber et al. 2010; Wagenknecht 2016; Fellows 2017). Our use of this concept will be anchored in two core assumptions. First, trust is a dependence relation between a person or organization and another person or organization. The first agent depends on the second one to do something important for it. An individual molecular phylogeneticist, for example, may rely on GenBank (Clark et al. 2016) to maintain an up-to-date collection of DNA sequences, because developing such a resource on her own would be cost prohibitive and redundant. Second, a relation of dependence is elevated to being one of trust when the first agent cannot control or validate the second agent's actions. This might be because the first agent lacks the knowledge or skills to perform the relevant task, or because it would be too costly to check. Historically, when phylogeneticists published individual gene sequences or alignments in their research articles, one could actually check 'by hand' whether a GenBank entry was correct (Strasser 2010, 2011). Nowadays, few biologists are able to validate 'directly' whether a genome sequence was correctly assembled from next-generation sequencing data.

Trusting someone means being in a position of dependence, but not necessarily that one is defenseless. Instead, individuals often rely on higher-order signals about an agent's reliability and competence, which need not hinge on knowing what that agent knows or being there to see what that agent actually does (Hardwig 1985; Fricker 2007; Carrier 2010). Talking to biologists who have worked with an aggregator's data in the past, for instance, is a common way to learn whether these data can be trusted as accurate and for what purposes. Thus the critical studies cited above are valuable in part because they move the experience of working with aggregated data into the public domain. Reading these articles is analogous to reading on-line reviews about a contractor's performance on past jobs.

However, it is arguably at least as important to know what an agent does when a task goes wrong, particularly for complex tasks like building a house or a higher-level biodiversity data aggregation service. Trust, then, is tied to more than just the intrinsic accuracy of the data; it also depends critically on how the trusted agent responds to negative outcomes – whether due to epistemic disagreements, honest errors, unanticipated difficulties, or negligence (Carrier 2010; de Cruz and de Smedt 2013; Winsberg et al. 2015). Trust will be especially low if there are stark discrepancies between (1) an agent's actual contribution to perceived data deficiencies and (2) the degree of responsibility that this agent acknowledges for that contribution. More succinctly,

bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

trust is lowered when there is a mismatch between "guilt" and "admission" – i.e., an accountability gap – in the context of data shortcomings. Transparency about ongoing progress or known problems is often crucial to cementing an ongoing relationship of trust, because it demonstrates that the trusted agent is willing to pay the cost of generating regular reports (Dourish 2001; Morris et al. 2013). Some aggregators are intentionally transparent about letting experts identify data deficiencies by allowing the flagging of records with perceived problems (Gries et al. 2014; Belbin and Williams 2016).

Matching Accountability to Responsibility

It is especially harmful if the trusted agent creates new problems in an area of central concern to those relying on it, while refusing to be held accountable for correcting them. We will argue that this is the case with the generation of synthetic taxonomic classifications and phylogenies – so called "backbones" or "trees of life" – that prevail in many biodiversity data aggregation networks (e.g. Bisby 2000; Bisby and Roskov 2010; Hinchcliff et al. 2015; Jong et al. 2015; Vandepitte et al. 2015; GBIF Secretariat 2016). High-level choices in designing these environments routinely lead to the creation of novel hierarchical syntheses that re-structure the occurrence record data in scientifically significant ways (Leonelli 2013). In doing so, aggregators also systematically compromise established conventions of sharing and recognizing taxonomic work. Taxonomic experts play a critical role in licensing the formation of high-quality biodiversity data packages. Systems of accountability that undermine or downplay this role are bound to lower both expert participation and trust in the aggregation process.

It is pivotal, then, that aggregators have embraced a design paradigm that requires one hierarchy (at a time) to organize all occurrence data. The production of a unitary hierarchy is governed by feasibility constraints, such as computational complexity costs and institutional limits on sharing access to information, rather than principles based in best systematic theory and practice. In addition, achieving a unitary hierarchy requires aggregators to eliminate taxonomic conflict between data input sources. This often results in a hierarchy that no longer corresponds to the view of any particular source: it becomes a synthesis nobody believes in. Biologists frequently regard the quality of these novel classification theories as deficient.

Responsibilities for issues with aggregator-created synthetic classifications are not rightfully owned by any data-providing source. In particular, a one-time fix (adjustment) of the synthesis to match one source classification fails to correct an underlying design flaw: to achieve trustworthy bodies of data for biodiversity research and conservation, we need to manage each body of data as a coherent whole, not just as an aggregate that is somehow supposed to maintain unity at any scale. The process of aggregation is effectively designed to disrupt local data formation and unity constraints for each source (Fig. 1).

Whenever aggregation leads to unintended and undesirable consequences at the system level, a sustainable solution is to shift the distribution of powers between providers and aggregators that is entrenched in the system's design. Since high-quality data packages frequently mirror competing and conflicting classifications, i.e., unsettled systematic research, this plurality must be accommodated in the design of biodiversity data integration.

Generation of Novel Systematic Syntheses

We use the term "systematic syntheses" as an encompassing term for different ways of

bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

organizing biodiversity into hierarchies, ranging from Linnaean taxonomies to phylogenetic trees. Achieving systematic synthesis is the explicit goal of aggregators such as GBIF (GBIF Secretariat 2016), which assembles its classification from more than 50 sources which are themselves unitary systems for particular subdomains of life (e.g. Bisby and Roskov 2010; Peters and McClennen 2015; WoRMS Editorial Board 2017).

The aim to achieve one natural hierarchy is perhaps as old as systematics, remains valid today, and need not be dissected here. Similarly, the broader feasibility and merit of generating backbones have been discussed before in the present context (e.g. Bisby 2000; Godfray 2002; Scoble 2004; Godfray et al. 2007). The main, likely uncontroversial conclusion that we carry over from these exchanges, is this: in all instances where alternative, lower-level input classifications (subtrees) are available, it is necessary to select one schema over the alternative(s) to create the synthesis. This process often involves input from socially sanctioned individuals and committees, or (increasingly) the use of computer algorithms that resolve conflicts according to programmed criteria (Page and Valiente 2005; Hinchcliffe et al. 2015; Döring et al. 2016; Redelings and Holder 2016; Rees and Cranston 2017).

Regardless of whether achieved by committee or algorithm, systematic syntheses typically have unique histories of creation, and – by virtue of preserving the choices made to resolve conflicts – may include or exclude information about how life on earth is structured. In accordance with Leonelli (2013), these syntheses constitute *novel* classification *theories*. "Some classificatory systems systematically and synthetically express, rather than simply affect, knowledge obtained through scientific research, and they do it in a way that (1) is unique, since such knowledge is not formalized anywhere else in the same way; (2) has huge influence on knowledge-making practices; and (3) enables experimenters to make sense of the results they obtain. […] Articulating knowledge that enables scientists to assess and value their results is an achievement that goes well beyond listing a set of commonly used assumptions as a basis for further inquiry. In the latter case, existing knowledge is applied to put a given set of items into some order; in the former, existing knowledge is transformed and developed so as to facilitate the conceptual analysis of data" (Leonelli 2013: 344–345).

To give one example, the 2016 version of the GBIF taxonomy (GBIF Secretariat 2016), which is largely but not fully congruent with Ruggiero et al. (2015), recognizes 99 phyla in eight kingdoms. Of these, two phyla are in the Archaea sec. GBIF Secretariat (2016) and 29 phyla are in the Bacteria sec. GBIF Secretariat 2016 (we use the "sec." – according to – to specify the source's name usage; see Franz et al. 2016a, 2016b). Meanwhile, Hug et al.'s (2016) "new view of the tree of life" recognizes 26 phyla of Archaea sec. Hug et al. (2016) and 92 phyla of Bacteria sec. Hug et al. (2016). While Hug et al.'s (2016) hierarchy is more transparently advertised as an outcome of systematic inference than the GBIF taxonomy (GBIF Secretariat 2016), both are relevantly novel and unique in terms of their systematic content. Using one over the other meets Leonelli's (2013) criteria, i.e., such choices will influence how knowledge is transformed and how data analyses are facilitated. In short, by generating systematic syntheses, many aggregators are makers of novel theories that drive how these data enable inferential outcomes.

To give another example, primate taxonomy is arguably now experiencing a period of *de*-stabilization, where the definitional boundaries of primate species remain subject to disagreement in the presence of increasing amounts of data and inference tools (Rylands and Mittermeier 2014; Franz et al. 2016b). The GBIF taxonomy has an influential function in this context, because newly indexed records whose identified species-level names are not endorsed by the hierarchy are 'matched' to it in one of two ways (Gaiji et al. 2013): either (1) names

bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

recognized as invalid versions (synonyms) of valid names are replaced by those, or (2) names that are not recognized at all at the species level are represented only the next available higher rank (e.g. genus), with a "null value" at the species rank.

The act of modulating the taxonomic identity of an occurrence record is a form of scientific arbitration that differs from a mere "enabling" of that record for aggregation and retrieval. The act may challenge and overrule the original judgment of taxonomic validity at the same nomenclatural rank, or may altogether reject and change that rank assignment. Even if this is not the case, transformation of the original taxonomic identities of records to match the chosen hierarchy constrains these records to reflect exactly those taxonomic judgments endorsed by the hierarchy (Fig. 1). Inferences of species distributions or threat status depend on such constraints (Peterson and Navarro-Sigüenza 1999; Rylands and Mittermeier 2014; Franz et al. 2016a).

Role of the Darwin Core Standard

The Darwin Core standard (Wieczorek et al. 2012) plays an underappreciated role in creating the accountability gap. Darwin Core competes in this context with another standard for exchanging biodiversity data: the Taxonomic Concept Transfer Schema (TCS; Kennedy et al. 2006). The TCS has a more limited scope than DwC with regards to non-taxonomic meta-/data properties of occurrence records. Yet the TCS was specifically designed to represent and integrate change and conflict across systematic hierarchies (e.g. Kennedy et al. 2006; Franz et al. 2008, 2016a, 2016b; Remsen 2016). One of the key TCS conventions is to manage "taxonomic concept labels", i.e., to consistently refer to name usages *according to* (sec.) a particular source. This allows the assembly of multiple coherent taxonomic hierarchies, each of which may assign incongruent meanings to the same or to overlapping sets of names, with one shared platform. The taxonomic concepts (theory regions) endorsed by each hierarchy can be articulated using the relationships of Region Connection Calculus (RCC–5), which are part of the TCS. Such an approach can yield logically consistent, multi-hierarchy reconciliation maps (alignments), without disrupting the perspective advocated by each data source (Lepage et al. 2014; Franz et al. 2016a, 2016b).

By processing biodiversity data primarily via DwC, aggregators buy into a model that fails to represent the sources' data signals directly. Because DwC lacks the syntactic conventions needed to absorb and align conflicting taxonomic hierarchies, the choice to favor DwC- over TCS-based solutions becomes a systemic weakness of the aggregation design.

Problems with relying just on DwC syntax are further amplified by widespread implementation practices. For instance, while DwC permits the use of taxonomic concept labels at the species level, the standard does not forbid leaving the "DwC: nameAccordingTo" category empty. Because enforcing such labels is optional in DwC, doing so now becomes a social responsibility. In practice, the most occurrence records that participate in global aggregation are syntactically under-specified at the species level. GBIF itself does not source species-level names to individual, expert-authored publications (GBIF Secretariat 2016).

Conversely, DwC does not require filling in higher-level taxonomic names (genus, etc.) for occurrence records. Yet in this case the option to do so *should* be ignored, because above the species-level DwC syntax does not permit using taxonomic concept labels at all. Hence the DwC-permissible higher-level name usages are necessarily under-specified. In either case, these higher-level names need not travel along with every occurrence record. The species-level taxonomic concept label is sufficient to indicate the record's identity with provenance from an

bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

expert-authored taxonomy. A full representation of that source's taxonomic signal, and of its conflicts with other published signals, should live outside of the occurrence record.

In summary, the basic design preference for DwC over TCS and the way in which the former is typically implemented, are choices that compromise the taxonomic coherence of biodiversity data packages being processed for aggregation.

Aggregation and Authorship

Some aggregators may disagree that systematic syntheses are novel theories, or challenge the significance that we ascribe to their creation under common and pragmatic conventions. Several recent (self-)assessments of aggregators do not focus on the issue of *backbones* as an important challenge they must own and overcome (e.g. Parr et al. 2011; Belbin et al. 2013; Faith et al. 2013; Gaiji et al. 2013; Sikes et al. 2016). Instead, we find the following position symptomatic (Belbin et al. 2013: 73; italics added for emphasis): "*Agencies* such as the ALA and GBIF *enable* observations *to be recorded directly* to their systems. These records are reviewed before being 'published', but the ALA and GBIF are *not* the data provider and therefore *cannot assume responsibility* for these records."

In contrast, we argue that by promoting backbones, aggregators act not just as data access faciltators but as data identity authors. GBIF in effect prohibits representation of occurrence records under taxonomic views that either conflict with the endorsed view, reflect a more granular view, or which are not yet recognized in the synthesis. It is not equitable to suggest that these issues should be dealt with "at the source", or that "data errors are best addressed through collaboration between all relevant agencies" (Belbin et al. 2013: 73).

Furthermore, aggregators frequently publish their syntheses under conventions that obscure individual expert authorship. This is not an all-or-nothing phenomenon; often there are linkages to individually authors or authored-attributed sources at lower-level nodes of the synthesis. Yet for instance, the 50+ sources for the GBIF taxonomy are all cited as initiatives, i.e., institutionalized catalogues, checklists, databases, species files projects, etc. No individual authors are named 'on the surface'. The entire synthesis is published by the "GBIF Secretariat" (2016). The Open Tree of Life project (Hinchliff et al. 2015; McTavish et al. 2015; Redelings and Holder 2016), which is groundbreaking in its ability to accommodate individual source publications, nevertheless publishes its periodical synthesis versions without naming individual authors. The World Register of Marine Species lists ca. 260 taxonomic editors, making an effort to accredit views to editors and primary publications at lower levels. That said, the "WoRMS Editorial Board" (2017) publishes the entire synthesis.

Publication conventions that obscure individual expert authorship of novel syntheses can impact trust. Such conventions are made feasible in part by the use of data standards – e.g. Darwin Core – that allow creating and publishing large, web-only syntheses from myriads of individual, traditionally accredited publications. Unlike just three decades ago, aggregators now have technical means to appropriate and synthesize thousands of expert-authored monographic, revisionary, and phylogenetic publications into novel syntheses that *on the surface* accredit only an initiative or committee. However, the social costs of creating syntheses with obscured author accreditation can be considerable. To expert authors, this may signal not only that some low-level form of intellectual appropriation is taking place (Wägele et al. 2011), but more disruptively, that accountability standards for the process of generating new systematic theories are shifting. For any particular taxonomic group, the synthesis may well represent the majority

bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

view. But who is responsible for defending that view scientifically? Who can be engaged to receive credit, and also criticism, for creating the synthesis as a unique hypothesis of how the natural world is structured?

Historically, the field of systematics has provided built-in opportunities for individual authors to reaffirm, expand, or challenge existing classifications through persistent publication venues. If aggregators publish syntheses that are not just novel but authorless in the conventional sense, they are thereby redesigning both the content of systematic theories and the social mechanisms for assigning responsibility for such content. Particularly the latter action constitutes a change in the power structure between aggregators and individual experts, to the experts' detriment.

Consequences of Disenfranchising Taxonomic Experts

Taxonomic perspectives that conflict with an aggregator's synthesis have no equitable way to compete in the same environment. Consider, for instance, the effect of this single-party system on the aggregators' relationships with (e.g.) early-career systematists. Many systematists will find their path into the field because of the following circumstances: they have a passion for understanding biodiversity and have likely had a formative experience that existing systematic hypotheses for a particular lineage are empirically inadequate. *Not* accepting the consensus is an important motivator for systematic careers. The mission to revise challenging groups solidifies the intellectual identity of systematists, who must become professional consensus disruptors to make their mark in science.

Systematic challenges that persist today are typically not trivial. Resolving them requires narrow specialization; hence the term "expert" is appropriate to identify an individual's record of training and accomplishment. The research products of early-career systematists, including monographic or revisionary treatments generated to meet thesis requirements, are usually published in international, peer-reviewed journals. Traditionally, this is how novel, high-quality hierarchies become part of the systematic knowledge base while advancing systematists' career trajectories.

It should be unproblematic for a graduate student's published monograph, including the therein newly identified occurrence records, to *immediately* be integrated with global aggregator environments – even and especially if the new classification conflicts with the aggregator's synthesis, as it almost invariably will. Technical and social barriers to adding these data *as published* have a strong antagonistic effect. Evidently the monograph was good enough to earn a doctoral degree and peer-reviewed publication – key accolades for reaching the next professional levels – but somehow, interfacing with the aggregator environment requires another level of validation that appears unattainable.

The example of the monograph of *Zelus* Fabricius, 1803 sec. Zhang et al. (2016) illustrates how the current power structure can operate. This publication accommodates 71 species-level concepts, including 24 new species names and numerous other taxonomic and nomenclatural changes. The monograph also identifies 10,628 specimen records. These can be downloaded in DwC-compatible format from the article website of the *Biodiversity Data Journal* (see also Smith et al. 2013). The example is fair precisely because the occurrence records were shared openly at the time of publication and in accordance with conventions that the aggregator endorses (GBIF 2017). This is not a case of erroneous execution.

Zhang et al. (2016) was published in early July, 2016. Later that month, the records were aggregated in GBIF as well. However, the aggregation process *newly* constrained the original set

bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

of occurrence records in at least two critical ways. First, the aggregator (GBIF 2017) shows only 409 of the source records (3.8%), i.e., those cited in the publication's main text. This constraint was applied by the journal publisher. Second, for the duration of nearly 220 days – i.e., until the next GBIF Backbone update took place (GBIF Developer Blog 2017) – the aggregator validated only 17 of 78 taxonomic names (21.8%), leaving 61 species-level epithets unrecognized. Among other consequences, this meant that the epithets of 40 holotype specimens referenced in Zhang et al. (2016) were changed to show only the genus-level name "*Zelus* Fabricius, 1802" in the GBIF aggregate.

The GBIF Backbone is not open for direct edits by expert authors who would validate the respective epithets sec. Zhang et al. (2016). Directly submitting (e.g.) a Darwin Core Archive data file (GBIF 2010) is also not an option for individual authors. Experts can only submit "issues" for tracking through the ChecklistBank code repository (Döring et al. 2017), with the expectation that someone with high-level access rights will address them (Mesibov 2013).

It is clearly positive that within seven months, the GBIF records' identifications were adjusted to reflect Zhang et al. (2016). At the same time, neither that event nor the entire process can be said to have empowered the authors of the monograph. The vibe that experts get from an aggregator is more like this: *we* choose if we want to represent your knowledge, and when. One might even say: if an aggregator wanted to 'hit' early-career systematists where it hurts their academic identities the most, controlling the form and content of their thesis-derived occurrence record identifications and systematic hypotheses is unfortunately quite effective. In analogy to the effect of negative on-line reviews for contractors, the experience of feeling disenfranchised will endure long after the records are adjusted.

Backbone-Based Data Signal Distortion: An Example

The insect monograph of Zhang et al. (2016) may come to enjoy community-wide acceptance for many years to come, effectively replacing all predecessors. Yet the classifications of other organismal groups are revised more frequently, with the outcome that multiple incongruent systems remain in use *simultaneously* (e.g. Peterson & Navarro-Sigüenza 1999; Lepage et al. 2014; Franz et al. 2016a, 2016b, 2016c). In some instances, the parallel existence of regional biodiversity data cultures that endorse conflicting taxonomies reflects an equilibrium condition (cf. Lepage et al. 2014). Persistent conflict is also common in phylogenomics, where competing tree hypotheses can appear in short order and continue to attract endorsements by third-party researchers (Brown et al. 2017).

Taxonomic backbones are deeply problematic in such situations. Our 'mostly real' example of Fig. 1 shows how a backbone may distort the identities of occurrence records *qua* aggregation, and thereby support inferences that conflict with every source-derived data signal. The example is a simplification of figures 1 and 2 in Franz et al. (2016a), showing only 20 occurrences that originate from the Southeast Regional Network of Expertise and Collections herbarium portal (SERNEC Data Portal 2017). These were deliberately filtered to illustrate the effect.

The starting conditions are common enough. A group of endangered orchids is subject to multiple taxonomic revisions – reviewed in Weakley (2015) – over a relatively short time interval. The revisions have led to a complex matrix of name usage relationships between different sources, with incongruent perspectives regarding species-level concept granularity and genus-level concept assignment. In spite of this, all expert-promoted views (A–C in Fig. 1) concur internally that two of the four ecoregions herein identified – i.e., R1 and R4 – do not

bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

harbor individuals from multiple recognized species in the group. Furthermore, no expert author has ever recognized the presence of an entity with the epithet *bifaria* in the ecoregion R2. That said, and depending on the taxonomies being aligned, the names *bifaria* and *divaricata* have varying, pro parte synonymy relationships.

Backbone-based aggregation fails to preserve the expert-sourced taxonomic coherence of the 20 occurrence records in our example (D and E in Fig. 1). The data-transforming process yields novel biological inferences with no support from any input source; viz. R1 and R4 are regions of sympatry for the orchid group, whereas region R2 harbors an entity labeled *bifaria* yet not *divaricata.* In other words, the interaction of taxonomic pluralism at the source level with backbone-based aggregation can create type I or type II errors in distributional signals (i.e., false positives or negatives) for which only aggregators may claim responsibility. Given the power balance inherent in the aggregation design, such errors need not be frequent to have a chilling effect on biologists' trust.

CONCLUSIONS

Our diagnosis implies a clear set of recommendations for moving forward. A first step is to recognize that trust is not just a feature of data signal quality but also a consequence of the social design of aggregation and the resulting power balance between data contributors and aggregators. This insight should not count as a justification for contributors to withhold occurrence records (cf. Ferro and Flick 2015; Sikes et al. 2016). However, not trusting an out-of-balance social design *does* make sense in light of well-established thinking about cooperative knowledge systems (Carrier 2010; Fricker 2007; Sperber et al. 2010; Winsberg et al. 2015; Wagenknecht 2016; Fellows 2017). Aggregators are well advised to accept an inclusive notion of trust that gives weight to equitable social engineering.

A second step is to acknowledge the accountability gap created by downplaying the impact of taxonomic backbones. Providing a clearinghouse for occurrence records is not simply a matter of cleaning up records provided by source collections. Instead, one needs a way to collate those records into meaningful biological datasets for swathes of life where no universal, consensus hierarchy exists. The practical need for such a hierarchy does not erase its scientific status as a classification theory (Leonelli 2013).

We suggest that aggregators must either author these classification theories in the same ways that experts author systematic monographs, or stop generating and imposing them onto incoming data sources. The former strategy is likely more viable in the short term, but the latter is the best long-term model for accrediting individual expert contributions. Instead of creating hierarchies they would rather not 'own' anyway, aggregators would merely provide services and incentives for ingesting, citing, and aligning expert-sourced taxonomies (Franz et al. 2016a).

As a social model, the notion of backbones (Bisby 2000) was misguided from the beginning. They disenfranchise systematists who are by necessity consensus-breakers, and distort the coherence of biodiversity data packages that reflect regionally endorsed taxonomic views. Henceforth, backbone-based designs should be regarded as an impediment to trustworthy aggregation, to be replaced as quickly and comprehensively as possible. We realize that just saying this will not make backbones disappear. However, accepting this conclusion counts as a step towards regaining accountability.

A third step is to refrain from defending backbones as the only pragmatic option for aggregators (Franz 2016). The default argument points to the vast scale of global aggregation

bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

while suggesting that only backbones can operate at that scale now. The argument appears valid on the surface, i.e., the scale *is* immense and resources are limited. Yet using scale as an obstacle it is only effective *if* experts were immediately (and unreasonably) demanding a fully functional, all data-encompassing alternative. If on the other hand experts are looking for *token actions* towards changing the social model, then an aggregator's pursuit of smaller-scale solutions is more important than succeeding with the 'moonshot'. And clearly, such solutions *have* been developed, often under the term "taxonomic concept approach" (e.g. Berendsohn 1995; Kennedy et al. 2006; Franz et al. 2008). Avibase (Lepage et al. 2014) uses this approach to manage more than 1.5 million taxonomic concept labels from 150+ checklists published over 125 years. This scale, covering some 10,000 avian species-level concepts as currently recognized by distinct sources, was achieved largely through an exceptional single-person effort. Aggregators seeking to improve expert trust are therefore better advised to embrace such design alternatives now rather than focus too much on global scalability as part of an invalid all-or-nothing argument. Otherwise, what is sold as pragmatism begins to sound dogmatic. In the presence of successful medium-sized solutions, the unwillingness to adjust designs – even for pilot projects – will likely be viewed by the expert community as a strategy to double down on the current power structure.

Another seemingly pragmatic argument is that many data packages to be aggregated are themselves syntactically underspecified. Again, while true on the surface, this does not preclude the design of pilot systems that enforce TCS- over DwC-based syntax – at a scale commensurate with individual expert publications (e.g. Senderov and Penev 2016). Failures to propagate such innovations at greater scales, and continued preferences for Darwin Core for purposes beyond its design scope, send the wrong message to experts.

Which brings us to the final conclusions. We intended to diagnose a systemic impediment to trusting biodiversity data aggregation. Whenever we mentioned specific aggregators, our purpose was not to engage in name-calling but to exemplify a broader design paradigm. While we cannot claim to have a full-scale solution on hand, the key directive is to develop new technical pathways and social incentives for experts to contribute directly to the validation of taxonomically coherent data packages in a greater aggregate. This will require new, broad-based political will in order to ensure its priority of the agendas of aggregators.

Over the past several decades, biodiversity data aggregation has taken a turn away from the agenda of promoting the careers of individual systematists. While recognizing the reasons aggregators had for taking this path, we suggest that the price of doing so, translated into declining trust, is higher than aggregators may have expected. The price is likely also higher than aggregators can afford in order to maintain long-term viability in the biodiversity data 'economy'.

We view this diagnosis as a call to action for both the systematics and the aggregator communities to reengage with each other. For instance, the leadership constellation and informatics research agenda of entities such as GBIF or Biodiversity Information Standards (TDWG 2017) should strongly coincide with the mission to promote early-stage systematist careers. That this is not the case now is unfortunate for aggregators, who are thereby losing credibility. It is also a failure of the systematics community to advocate effectively for its role in the biodiversity informatics domain. Shifting the power balance back to experts is therefore a shared interest.

bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

FUNDING

This research was supported by the National Science Foundation, under the grants DEB–1155984, DBI–1342595 (NMF), and SES–1153114 (BWS).

ACKNOWLEDGMENTS

The authors are grateful to Erin Barringer-Sterner, Andrew Johnston, Jonathan Rees, David Remsen, David Shorthouse, and Guanyang Zhang for helpful discussions on this subject.

REFERENCES

Baker, B. 2011. New push to bring U.S. biological collections to the world's online community. BioScience 61:657–662.

Ballesteros-Mejia L., Kitching I.J., Jetz W., Nagel P., Beck J. 2013. Mapping the biodiversity of tropical insects: species richness and inventory completeness of African sphingid moths. Glob. Ecol. Biogeogr. 22:586–595.

Beck J., Ballesteros-Mejia L., Nagel P., Kitching, I.J. 2013. Online solutions and the 'Wallacean Shortfall': what does GBIF contribute to our knowledge of species' ranges? Diversity Distrib. 19:1043–1050.

Belbin L., Daly J., Hirsch T., Hobern D., LaSalle J. 2013. A specialist's audit of aggregated occurrence records: an 'aggregator's' perspective. ZooKeys 305:67–76.

Belbin, L., Williams, K.J. 2016. Towards a national bio-environmental data facility: experiences from the Atlas of Living Australia. Int. J. Geogr. Inf. Sci. 30:108–125.

Berendsohn W.G. 1995. The concept of "potential taxa" in databases. Taxon 44:207–212.

Berg M. 1998. The politics of technology: on bringing social theory into technological design. Sci. Technol. Human Values 23:456–490.

Berman J.J. 2013. Principles of big data. Elsevier, Waltham, Massachusetts.

Bisby F.A. 2000. The quiet revolution: biodiversity informatics and the internet. Science 289:2309–2312.

Bisby F.A., Roskov Y.R. 2010. The Catalogue of Life: towards an integrative taxonomic backbone for biodiversity. In: Nimis P.L., Vignes Lebbe R., editors. Tools for identifying biodiversity: progress and problems. Proceedings of the International Congress, Paris, September 20–22, 2010. Edizioni Università di Trieste, Trieste, pp. 37–42.

Blagoderov V., Kitching I.J., Livermore L., Simonsen T.J., Smith V.S. 2012. No specimen left behind: industrial scale digitization of natural history collections. ZooKeys 209:133–146.

Bortolus A. 2008. Error cascades in the biological sciences: the unwanted consequences of using bad taxonomy in ecology. AMBIO 37:114–118.

Bowker, G.C. 2000. Biodiversity datadiversity. Soc. Stud. Sci. 30:643–683.

Brown, J.W., Wang N., Smith S.A. 2017. The development of scientific consensus: analyzing conflict and concordance among avian phylogenies. bioRxiv. Available from https://doi.org/10.1101/123034 Accessed 01 June 2017.

Carrier M. 2010. Scientific knowledge and scientific expertise: epistemic and social conditions of their trustworthiness. Analyse & Kritik 32:195–210.

Clark K., Karsch-Mizrachi I., Lipman D.J., Ostell J., Sayers E.W. 2016. GenBank. Nucleic Acids Res. 44:D67–D72.

bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

Costello M.J., Michener W.K., Gahegan M., Zhang Z.-Q., Bourne P., Chavan V. 2012. Quality assurance and intellectual property rights in advancing biodiversity data publications, Version 1.0. Copenhagen, Global Biodiversity Information Facility, pp. 1–33. Available from http://www.gbif.org/resource/80818 accessed 01 June 2017.

Datta A., Sen S., Zick Y. 2016. Algorithmic transparency via quantitative input influence: theory and experiments with learning systems. 016 IEEE Symposium on Security and Privacy; pp. 598–617.

De Cruz H., de Smedt J. 2013. The value of epistemic disagreement in scientific practice. The case of Homo floresiensis. Stud. Hist. Phil. Sci. 44:169–177.

Döring M., Méndez Hernández F., et al. 2017. GBIF Checklistbank. Available from https://github.com/gbif/checklistbank accessed 01 June 2017.

Dourish P. 2001. Process descriptions as organisational accounting devices: the dual use of workflow technologies. Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work; pp. 52–60.

Edwards J.L. 2004. Research and societal benefits of the Global Biodiversity Information Facility. Bioscience 54: 485–486.

Edwards P.N., Mayernik M.S., Batcheller A.L., Bowker G.C., Borgman C.L. 2011. Science friction: data, metadata, and collaboration. Soc. Stud. Sci. 41:667–690.

Faith D.P., Collen B., Ariño A.H., Koleff P., Guinotte J., Kerr J., Chavan V. 2013. Bridging the biodiversity data gaps: recommendations to meet users' data needs. Biodiv. Inform. 8:41–58.

Fellows J.J. 2017. Trust without shared belief: pluralist realism and polar bear conservation. Perspect. Sci. 25:36–66.

Fernald M.L. 1950. Gray's manual of botany, eighth (centennial) edition. American Book Company, New York.

Ferro M.L., Flick A.J. 2015. "Collection bias" and the importance of natural history collections in species habitat modeling: a case study using *Thoracophorus costalis* Erichson (Coleoptera: Staphylinidae: Osoriinae), with a critique of GBIF.org. Coleop. Bull. 69:415–425.

Franklin J., Serra-Diaz J.A., Syphard A.D., Regan H.M. 2017. Big data for forecasting the impacts of global change on plant communities. Global Ecol. Biogeogr. 26:6–17.

Franz, N.M., editor. 2016. "Who authors GBIF's Backbone?" Available from https://storify.com/taxonbytes/who-authors-gbif-s-backbone accessed 01 June 2017.

Franz N.M., Chen M., Yu S., Kianmajd P., Bowers S., Weakley, A.S., Ludäscher B. 2016c. Names are not good enough: reasoning over taxonomic change in the *Andropogon* complex. Semantic Web (IOS) 7:645 –667.

Franz N., Gilbert E., Ludäscher B., Weakley A. 2016a. Controlling the taxonomic variable: taxonomic concept resolution for a southeastern United States herbarium portal. Res. Ideas Outcomes 2:e10610.

Franz N.M., Peet R.K., Weakley A.S. 2008. On the use of taxonomic concepts in support of biodiversity research and taxonomy. In: Wheeler Q.D., editor. The new taxonomy. Systematics Association Special Volume Series, Volume 74. Taylor & Francis, Boca Raton, pp. 63–86.

Franz N.M., Pier N.M., Reeder D.M., Chen M., Yu S., Kianmajd P., Bowers S., Ludäscher B. 2016b. Two influential primate classifications logically aligned. Syst. Biol. 65:561–582.

Franz N.M., Thau D. 2010. Biological taxonomy and ontology development: scope and limitations. Biodiv. Inform. 7:45–66.

Fricker M. 2007. Epistemic injustice: power and the ethics of knowing. Oxford University Press,

bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

New York.
Gaiji S., Chavan V., Ariño A.H., Otegui J., Hobern D., Sood R., Robles E. 2013. Content assessment of the primary biodiversity data published through GBIF network: status, challenges and potentials. Biodivers. Inform. 8:94–172.
García-Roselló E., Guisande C., Manjarrés-Hernández A., González-Dacosta J., Heine J., Pelayo-Villamil P., González-Vilas L., Vari R.P., Vaamonde A., Granado-Lorencio C., Lobo J.M. 2015. Can we derive macroecological patterns from primary Global Biodiversity Information Facility data? Global Ecol. Biogeogr. 24:335–347.
GBIF. 2010. Darwin Core Archives – How-to guide, version 1, released on 01 March 2011. Contributed by Remsen D., Braak K., Döring M., Robertson, T. Global Biodiversity Information Facility, Copenhagen, pp. 1–21. Available from http://links.gbif.org/gbif_dwca_how_to_guide_v1 accessed 01 June 2017.
GBIF. 2017. GBIF.org (12th February 2017). GBIF Occurrence Download. DatasetKey: A taxonomic monograph of the assassin bug genus *Zelus* Fabricius (Hemiptera: Reduviidae): 71 species based on 10,000 specimens. Records included: 409 records from 1 published datasets. DOI: http://doi.org/10.15468/dl.zhyqxp Available from http://www.gbif.org/occurrence/download/0059534-160910150852091 accessed 01 June 2017.
GBIF Developer Blog. 2017. GBIF Backbone – February 2017 update. Available from http://gbif.blogspot.com/2017/02/gbif-backbone-february-2017-update.html accessed 01 June 2017.
GBIF Secretariat. 2016. GBIF backbone taxonomy. DOI: 10.15468/39omei. Available from http://www.gbif.org/dataset/d7dddbf4-2cf0-4f39-9b2a-bb099caae36c accessed 01 June 2017.
Godfray H.C.J. 2002. Challenges for taxonomy. Nature 417:17–19.
Godfray H.C.J., Clark B.R., Kitching I.J., Mayo S.J., Scoble M.J. 2007. The web and the structure of Taxonomy. Syst. Biol. 56:943–955.
Graham C.H., Ferrier S., Huettman F., Moritz C., Peterson A.T. 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends Ecol. Evol. 19:497–503.
Gries C., Gilbert E.E, Franz N.M. 2014. Symbiota – a virtual platform for creating voucher-based biodiversity information communities. Biodiv. Data J. 2:e1114.
Gueta T., Carmel Y. 2016. Quantifying the value of user-level data cleaning for big data: a case study using mammal distribution models. Ecol. Inform. 34:139–145.
Harding A., Tolleya K.A., Burgessa N.D. 2015. Red List assessments of East African chameleons: a case study of why we need experts. Oryx 49:652–658.
Hardwig J. 1985. Epistemic dependence. J. Phil. 82:335–349.
Hardwig J. 1991. The role of trust in knowledge. J. Phil. 88:693–708.
Hill A.W., Otegui J., Ariño A.H., Guralnick R.P. 2010. GBIF position paper on future directions and recommendations for enhancing fitness-for-use across the GBIF network, Version 1.0. Copenhagen, Global Biodiversity Information Facility, pp. 1–25. Available from http://www.gbif.org/resource/80623 accessed 01 June 2017.
Hinchcliff C.E., Smith S.A., Allman J.F., Burleigh G., Chaudhary R., Coghill L.M., Crandall K.A., Deng J., Drew B.T., Gazis R., Gude K., Hibbett D.S., Katz L.A., Laughinghouse H.D. IV, McTavish E.J., Midford P.E., Owen C.L., Ree R.H., Rees J.A., Soltis D.E., Williams T., Cranston K.A. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc. Natl. Acad. Sci. 112:12764–12769.

bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

Hug L.A., Baker B.J., Anantharaman K., Brown C.T., Probst A.J., Castelle C.J., Butterfield C.N., Hernsdorf A.W., Amano Y., Ise K., Suzuki Y., Dudek N., Relman D.A., Finstad K.M., Amundson R., Thomas B.C., Banfield J.F. 2016. A new view of the tree of life. Nat. Microbiol. 1:16048.

Jetz W., McPherson J.M., Guralnick R.P. 2012. Integrating biodiversity distribution knowledge: toward a global map of life. Trends Ecol. Evol. 27:151–159.

Jong Y. de., Kouwenberg J., Boumans L., et al. (111 additional co-authors). 2015. PESI – a taxonomic backbone for Europe. Biodiv. Data J. 3:e5848.

Kartesz J. 2010. Floristic synthesis of North America, version 9-15-2010. Biota of North America Program (BONAP), Chapel Hill. Available from http://www.bonap.org/ accessed June 01 2017.

Kennedy J., Hyam R., Kukla R., Paterson T. 2006. Standard data model representation for taxonomic information. OMICS 10:220–230.

Leonelli S. 2013. Classificatory theory in biology. Biol. Theory 7:338–345.

Leonelli S. 2016. Data-centric biology: a philosophical study. University of Chicago Press, Chicago.

Lepage D., Vaidya G., Guralnick R. 2014. Avibase – a database system for managing and organizing taxonomic concepts. ZooKeys 420:117–135.

Maldonado C., Molina C.I., Zizka A., Persson C., Taylor C.M., Albán J., Chilquillo E., Rønsted N., Antonelli A. 2015. Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? Global Ecol. Biogeogr. 24:973–984.

Mcdowell A. 2002. Trust and information: the role of trust in the social epistemology of information science. Soc. Epistemol. 16:51–63.

McTavish E.J., Hinchliff C.E., Allman J.F., Brown J.W., Cranston K.A., Holder M.T., Rees J.A., Smith S.A. 2015. Phylesystem: a gitbased data store for community-curated phylogenetic estimates. Bioinformatics 31:2794–2800.

Mesibov, R. 2013. A specialist's audit of aggregated occurrence records. ZooKeys 293:1–18.

Meyer C., Kreft H., Guralnick R., Jetz W. 2015. Global priorities for an effective information basis of biodiversity distributions. Nat. Commun. 6:8221.

Millerand F., Ribes D., Baker K.S., Bowker G.C. 2013. Making an issue out of a standard: storytelling practices in a scientific community. Sci. Technol. Human Values 38:7–43.

Morris R.A., Dou L., Hanken J., Kelly M., Lowery D.B., Ludäscher B., Macklin J.A., Morris P.J. 2013. Semantic annotation of mutable data. PLoS ONE 8(11):e76093.

O'Malley M.A. 2013. When integration fails: prokaryote phylogeny and the tree of life. Stud. Hist. Philos. Biol. Biomed. Sci. 44:551–562.

Otegui J., Ariño A.H., Chavan V., Gaiji S. 2013a. On the dates of the GBIF mobilized primary biodiversity data records. Biodiv. Inform. 8:173–184.

Otegui J., Ariño A.H., Encinas M.A., Pando F. 2013b. Assessing the primary data hosted by the Spanish node of the Global Biodiversity Information Facility (GBIF). PLoS ONE 8(1):e55144.

Page R.D.M. 2008. Biodiversity informatics: the challenge of linking data and the role of shared identifiers. Brief. Bioinform. 9:345–354.

Page R.D.M., Valiente G. 2005. An edit script for taxonomic classifications. BMC Bioinformatics 6(1):208.

Parr C.S., Guralnick R., Cellinese N., Page R.D.M. 2011. Evolutionary informatics: unifying knowledge about the diversity of life. Trends Ecol. Evol. 27:94–103.

Peters S.E., McClennen M. 2015. The Paleobiology Database application programming

bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

interface. Paleobiology 42:1–7.
Peterson A.T., Soberón J., Krishtalka L. 2015. A global perspective on decadal challenges and priorities in biodiversity informatics. BMC Ecol. 15:15.
Peterson A.T., Navarro-Sigüenza A. 1999. Alternate species concepts as bases for determining priority conservation areas. Conserv. Biol. 13:427–431.
Por F.D. 2007. A "taxonomic affidavit": why it is needed? Integr. Zool. 2:57–59.
Radford A.E., Ahles H.E., Bell C.R. 1968. Manual of the vascular flora of the Carolinas. University of North Carolina Press, Chapel Hill.
Redelings B.D., Holder M.T. 2016. A supertree pipeline for summarizing phylogenetic and taxonomic information for millions of species. PeerJ Preprints 4:e2538v1. Available from https://doi.org/10.7287/peerj.preprints.2538v1 accessed 01 June 2017.
Rees J.A., Cranston K. 2017. Automated assembly of a reference taxonomy for phylogenetic data synthesis. Biodiv. Data J. 5:e12581.
Remsen D. 2016. The use and limits of scientific names in biological informatics. In: Michel E., editor. Anchoring biodiversity information from Sherborn to the 21st century and beyond. ZooKeys 550:207–233.
Ruggiero M.A., Gordon D.P., Orrell T.M., Bailly N., Bourgoin T., Brusca R.C., Cavalier-Smith, T., Guiry M.D., Kirk P.M. 2015. A higher level classification of all living organisms. PLoS ONE 10(6):e0130114.
Rylands A.B., Mittermeier R.A. 2014. Primate taxonomy: species and conservation. Evol. Anthr. 23:8–10.
Scoble M.J. 2004. Unitary or unified taxonomy? Philos. Trans. R. Soc. Lond. B 359:699–710.
Senderov V., Penev L. 2016. The Open Biodiversity Knowledge Management System in scholarly publishing. Res. Ideas Outcomes 2:e7757.
SERNEC Data Portal. 2017. Available from http://sernecportal.org Accessed 01 June 2017.
Sikes D.S., Copas K., Hirsch T., Longino J.T., Schigel D. 2016. On natural history collections, digitized and not: a response to Ferro and Flick. ZooKeys 618:145–158.
Simon J. 2010. A socio-epistemological framework for scientific publishing, Soc. Epistemol. 24:201–218.
Smith B.E., Johnston M.K., Lücking R. 2016. From GenBank to GBIF: phylogeny-based predictive niche modeling tests accuracy of taxonomic identifications in large occurrence data repositories. PLoS ONE 11(3):e0151232.
Smith V., Georgiev T., Stoev P., Biserkov J., Miller J., Livermore L., Baker E., Mietchen D., Couvreur T.L.P., Mueller G., Dikow T., Helgen K.M., Frank J., Agosti D., Roberts D., Penev L. 2013. Beyond dead trees: integrating the scientific process in the Biodiversity Data Journal. Biodivers. Data J. 1:e995.
Sperber D., Clément F., Heintz C., Mascaro O., Mercier H., Origgi G., Wilson D. 2010. Epistemic vigilance. Mind & Lang. 25:359–393.
Soberón J., Arriaga L., Lara L. 2002. Issues of quality control in large, mixed-origin entomological databases. In: Saarenmaa H., Nielsen E., editors. Towards a global biological information infrastructure, Volume 70. European Environment Agency, Copenhagen, pp. 15–22.
Sterner B.W., Franz N.M. 2017. Cognitive pragmatics for big biodiversity data: taxonomy for humans or computers? Biol. Theory. 12:99–111.
Strasser B.J. 2010. Collecting, comparing, and computing sequences: the making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954-1965. J. Hist. Biol. 43: 623–660.

bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

Strasser B.J. 2011. The experimenter's museum: GenBank, natural history, and the moral economies of biomedicine. Isis 102:60–96.

Stropp J., Lade R.J, Malhado A.C.M., Hortal J., Gaffuri J., Temperley W.H., Skøien J.O, Mayaux P. 2016. Mapping ignorance: 300 years of collecting flowering plants in Africa. Global Ecol. Biogeogr. 25:1085–1096.

USDA Plants. 2012. The PLANTS Database. National Plant Data Team, Greensboro. Available from http://plants.usda.gov Accessed 01 June 2017.

Turel O., Gefen D. 2015. The dual role of trust in system use. J. Comput. Inform. Syst. 54:2–10.

Vandepitte L., Vanhoorne B., Decock W., Dekeyzer S., Verbeeck A.T., Bovit L., Hernandez F., Mees J. 2015. How Aphia – the platform behind several online and taxonomically oriented databases – can serve both the taxonomic community and the field of biodiversity informatics. J. Mar. Sci. Eng. 3:1448–1473.

Yesson C., Brewer P.W., Sutton T., Caithness N., Pahwa J.S., Burgess M., Gray W.A., White R.J., Jones A.C., Bisby F.A., Culham A. 2007. How global is the global biodiversity information facility? PLoS One 2(11):e1124.

Wägele H., Klussmann-Kolb A., Kuhlmann M., Haszprunar G., Lindberg D., Koch A., Wägele J.W. 2011. The taxonomist – an endangered race. A practical proposal for its survival. Front. Zool. 8:25.

Wagenknecht, S. 2016. A social epistemology of research groups. Palgrave Macmillan, London.

Weakley A.S. 2015. Flora of the Southern and Mid-Atlantic States. University of North Carolina Herbarium, Chapel Hill. Available from http://www.herbarium.unc.edu/flora.htm accessed 01 June 2017.

Wheeler Q.D., Knapp S., Stevenson D.W., Stevenson J., Blum S.D., Boom B.M., Borisy G.G., Buizer J.L., De Carvalho M.R., Cibrian A., Donoghue M.J., Doyle V., Gerson E.M., Graham C.H., Graves P., Graves S.J., Guralnick R.P., Hamilton A.L., Hanken J., Law W., Lipscomb D.L., Lovejoy T.E., Miller H., Miller J.S., Naeem S., Novacek M.J., Page L.M., Platnick N.I., Porter-Morgan H., Raven P.H., Solis M.A., Valdecasas A.G., Van Der Leeuw S., Vasco A., Vermeulen N., Vogel J., Walls R.L., Wilson E.O., Woolley J.B. 2012. Mapping the biosphere: exploring species to understand the origin, organization and sustainability of biodiversity. Syst. Biodiv. 10:1–20.

Wieczorek J., Bloom D., Guralnick R., Blum S., Döring M., Giovanni R., Robertson, Vieglais D. 2012. Darwin Core: an evolving community-developed biodiversity data standard. PLoS ONE 7(1):e29715.

Winsberg E., Huebner B., Kukla R. 2015. Accountability and values in radically collaborative research. Stud. Hist. Phil. Sci. 46:16–23.

Wiser S.K. 2016. Achievements and challenges in the integration, reuse and synthesis of vegetation plot data. J. Veg. Sci. 27:868–879.

WoRMS Editorial Board. 2017. World Register of Marine Species. DOI: 10.14284/170. Available from http://www.marinespecies.org accessed 01 June 2017.

Zermoglio P.F., Guralnick R.P., Wieczorek J.R. 2016. A standardized reference data set for vertebrate taxon name resolution. PLoS ONE 11(1):e0146894.

Zhang G., Hart E.R., Weirauch C. 2016. A taxonomic monograph of the assassin bug genus *Zelus* Fabricius (Hemiptera: Reduviidae): 71 species based on 10,000 specimens. Biodiv. Data J. 4:e8150.

bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

FIGURE CAPTIONS

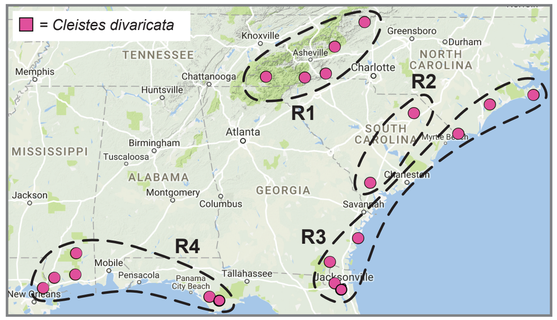
FIGURE 1. Backbone-based aggregation disrupts coherent biodiversity data packages. 'Most real' example adopted from Franz et al. (2016a). The top right table presents an alignment of five different taxonomies for the *Cleistes/Cleistesiopsis* complex sec. Radford et al. (1968), Fernald (1950), USDA Plants (2012), Kartesz (2010), and Weakley (2015). Columns indicate the relative congruence between different taxonomic concepts, whereas rows show the period of usage, validly recognized names, and sources. A–E: five representations of the same set of 20 specimens provided by the SERNEC Data Portal (2017), with distribution maps that identify four ecoregions R1–R4 (right), and tables displaying the ecoregion-specific presence (+), absence (–), or inapplicability (o – i.e., name not available) of occurrences identified to taxonomic concept labels. A–C: concept occurrence patterns according three reciprocally incongruent, yet internally coherent taxonomies; D: raw (unprocessed) aggregate of A–C, where each source contributes a complementary subset (data package) of the 20 specimens – hence six taxonomic names are shown; E: backbone-based transformation of D. Both D and E support new biological inferences (red circles) regarding the sympatry of multiple entities of the complex in ecoregions R1 and R4 (= false positives), and the local endemism of an entity labeled *bifaria* in R2 (= false negative), which is possible if pro parte synonymy relationships are not coherently transposed in the backbone-based synthesis.

Alignment of 5 taxonomic schemata (1922 – present)

#	Period of use	Concept lineage 1	Concept lineage 2	Concept lineage 3	According to (sec.)
5	2009 – present	*Cleistesiopsis divaricata*	*Cleistesiopsis oricamporum*	*Cleistesiopsis bifaria*	Weakley (2015)
4	2008 – present	*Cleistesiopsis divaricata*	*Cleistesiopsis bifaria*		Kartesz (2010)
3	1993 – present	*Cleistes divaricata*	*Cleistes bifaria*		USDA Plants (2012)
2	1946 – 1993	*Cleistes div.* var. *divaricata*	*Cleistes divaricata* var. *bifaria*		Fernald (1950)
1	1922 – 1991	*Cleistes divaricata*			Radford AB (1968)

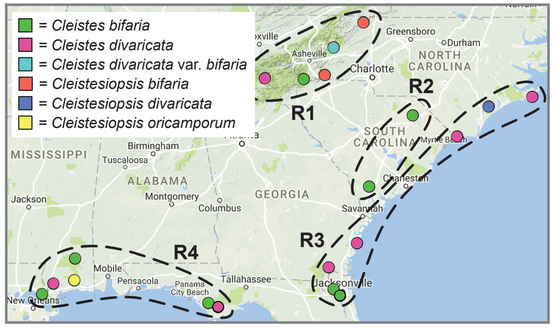
A. sec. Radford, Ahles & Bell (1968)

Distribution	*"bifaria"*	*"divaric."*	*"oricamp."*
R1	o	+	o
R2	o	+	o
R3	o	+	o
R4	o	+	o



Data package 1
=> Ingestion

D. sec. SERNEC (2016) - RAW AGGREGATION

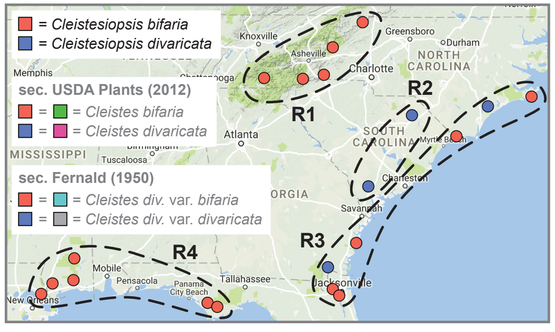


Distribution	*"bifaria"*	*"divaric."*	*"oricamp."*
R1	⊕	⊕	—
R2	⊕	⊖	—
R3	+	+	—
R4	⊕	⊕	⊕

bioRxiv preprint first posted online Jun. 28, 2017; doi: http://dx.doi.org/10.1101/157214. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

B. sec. Kartesz (2010) [BONAP]

Distribution	*bifaria*	*divaric.*	*oricamp.*
R1	+	—	o
R2	—	+	o
R3	+	+	o
R4	+	—	o



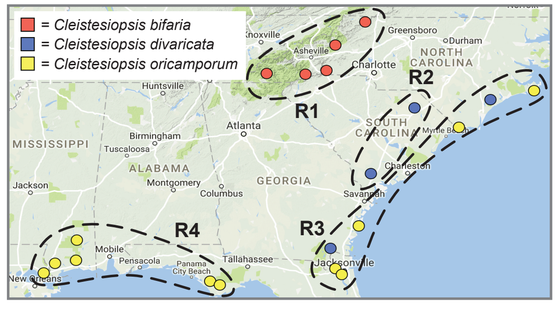
Data package 2
=> Ingestion

Data transformation to conform with single taxonomic backbone

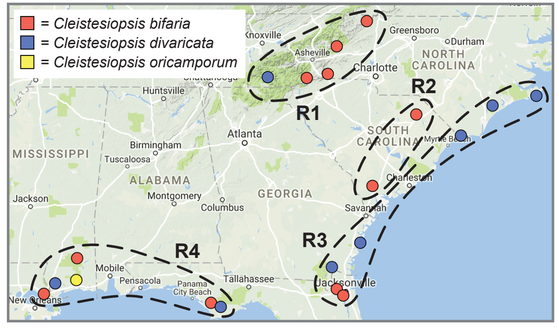
Aggregation yields *novel inferences* of sympatry (R1,R4) & endemism (R2)

C. sec. Weakley (2015)

Distribution	*bifaria*	*divaric.*	*oricamp.*
R1	+	—	—
R2	—	+	—
R3	—	+	+
R4	—	—	+



Data package 3
=> Ingestion



Distribution	*bifaria*	*divaric.*	*oricamp.*
R1	⊕	⊕	—
R2	⊕	⊖	—
R3	+	+	—
R4	⊕	⊕	⊕

E. sec. SERNEC (2016) - AGGREGATOR SYNTHESIS