Biol Theory (2017) 12:99–111
DOI 10.1007/s13752-017-0259-5

ORIGINAL ARTICLE

Taxonomy for Humans or Computers? Cognitive Pragmatics for Big Data

Beckett Sterner[1] · **Nico M. Franz**[1]

Received: 27 June 2016 / Accepted: 2 January 2017 / Published online: 15 February 2017

Abstract Criticism of big data has focused on showing that more is not necessarily better, in the sense that data may lose their value when taken out of context and aggregated together. The next step is to incorporate an awareness of pitfalls for aggregation into the design of data infrastructure and institutions. A common strategy minimizes aggregation errors by increasing the precision of our conventions for identifying and classifying data. As a counterpoint, we argue that there are pragmatic trade-offs between precision and ambiguity that are key to designing effective solutions for generating big data about biodiversity. We focus on the importance of theory-dependence as a source of ambiguity in taxonomic nomenclature and hence a persistent challenge for implementing a single, long-term solution to storing and accessing meaningful sets of biological specimens. We argue that ambiguity does have a positive role to play in scientific progress as a tool for efficiently symbolizing multiple aspects of taxa and mediating between conflicting hypotheses about their nature. Pursuing a deeper understanding of the trade-offs and synthesis of precision and ambiguity as virtues of scientific language and communication systems then offers a productive next step for realizing sound, big biodiversity data services.

Keywords Big data · Cognitive pragmatics · Concept taxonomy · Data aggregation · Knowledge representation and reasoning · Nomenclature

Introduction

Advocates for big data argue that more is different: as the amount and complexity of data available to a field increase by several orders of magnitude, radical changes become appropriate in scientific methodology (Anderson 2008; Hey et al. 2009; Mayer-Schönberger and Cukier 2013). The slogan suggests that entering a new era of data-driven science is primarily a matter of scaling up the quantity of data available to a field: how one gets the data matters less than how much one collects. In contrast, critics have argued that the value of an aggregated dataset cannot be reliably separated from how its component parts were produced (boyd and Crawford 2012; Lazer et al. 2014; Leonelli 2014; Meng 2014).[1] In particular, the process of aggregating data at increasing scale contains common pitfalls at various stages that can undermine the robustness of the data's value across contexts of reuse. In other words, each dataset undergoes its own journey from production to reuse (Leonelli 2016), and the details of these journeys matter for the outcome of the larger transformation the field undergoes as scientists scale up their data.

Given that process matters, how should scientists design their data infrastructures and practices in order to anticipate and overcome pitfalls in aggregation? While critics of the big data movement have argued for the difficulty of making data travel meaningfully across situations, we are still in need of systematic, positive alternatives to the naïve hope

✉ Beckett Sterner
beckett.sterner@asu.edu

[1] School of Life Sciences, Arizona State University, Tempe, AZ, USA

[1] We use "data aggregation" to refer to merging multiple sets of data of the same kind (e.g., multiple collections of specimens or multiple runs of the same experiment) as distinct from "data integration," which refers to combining multiple kinds of data to solve an inference problem (Berman 2013). The limits of this distinction, where aggregation and integration become hard to tell apart, are an important topic outside the scope of this article.

that adding more data will make any problem go away. For example, when is it better to deal with aggregation errors by developing a more precise, universal system of identifiers for data versus multiple systems specialized to different needs that are only partially consistent? We argue that there are pragmatic trade-offs between precision and ambiguity in the underlying syntax of identifier schemes that have important consequences for the design of solutions to aggregating biodiversity data in systematic biology. We introduce a framework for articulating these trade-offs that draws on recent research in cognitive pragmatics (Piantadosi et al. 2012).

In general, the theory-dependence of scientific data poses a persistent problem for a single solution to storing and accessing meaningful sets of scientific data. In this article we focus specifically on the problem for data semantics posed by the project of aggregating datasets that describe biological specimens preserved in collections across the world. When taxonomists aggregate primary specimen data into supposedly coherent evolutionary entities using taxonomic names, they rely on hypotheses about the nature of these entities that may conflict across specimen collections. That is, even when taxonomists recognize the same valid name string (or synonymy relationship between names), they regularly subscribe to different hypotheses about the evolutionary identity and definitional boundaries of the name's referent, which means they also often disagree about how to categorize new specimens or other data under existing taxonomic names. To put it another way, they associate conflicting *taxonomic concepts* with the same name, which means they disagree over what gets included in the taxon as a set of organisms.[2] The semantics of specimen data is therefore theory-dependent on taxonomic classifications, which themselves regularly conflict across research groups and time (Franz et al. 2015, 2016a, b; Witteveen 2015a; Remsen 2016).

Having recognized the theory-dependence of specimen data as a source of aggregation error, one response may be that biodiversity data is not a good candidate for big data projects. The contextuality and instability of taxonomic nomenclature simply demands too much human curation of data for scaling up to be worthwhile. Nevertheless, there are imperative reasons for building comprehensive biodiversity data environments, such as the need to establish an information system reflecting current perspectives on the geographic ranges of species in order to track the effects of climate change.

Indeed, big biodiversity data is effectively already here, emerging out of three major sources: the digitization of likely three billion specimens housed in museum collections and other institutions (Rogers 2016), the digitization of published monographs and texts from the past several hundred years that provide or revise taxonomic classifications (Page 2016), and the industrial-scale production of new data associated from other fields of biology, such as genomics. Rapid and affordable sequencing has enabled phylogenetic systematists to pursue building a universal tree of life (Hinchcliff et al. 2015), but reaching this goal depends on extensive prior knowledge about the lower-level entities that must be sampled and analyzed. There would be clear value to having all these forms of data linked together and made easily available from on-line databases, but this poses an immense challenge for biodiversity informatics.

We argue for a broader understanding of big data that focuses attention on its long-term outcomes across a diverse array of cases. To this end, we introduce the concept of a "big data trajectory" as the process by which a field changes the collective set of available data sufficiently to motivate major changes and new problems in existing research. A big data trajectory generally involves aggregating the results of multiple, and sometimes many, individual data journeys. From this broader perspective, the theory-dependence of data becomes a central factor in the success or failure of any community attempting to scale up its data.

There are currently two major strategies for addressing how to manage specimen data across multiple classification systems (Remsen 2016): syntactic (extending the internal grammar of nomenclature) and pragmatic (extending our knowledge about the modes and contexts of people's use of nomenclature). Some taxonomists, for example, have advocated for extending the syntax of taxonomic names to include a reference to an authoritative publication that explicitly defines the taxon concept being invoked (Berendsohn 1995). This solution can be directly implemented into current and future databases so that users are required to supply an explicit taxon concept label in order to enter new data. Several taxonomic publications, software projects, and databases have adopted Berendsohn's (1995) expanded syntax and implemented it to identify and track taxonomic concepts (e.g., Koperski et al. 2000; Pullan et al. 2000, 2005; Lepage et al. 2014; Franz et al. 2015, 2016a, b; Jansen and Franz 2015; Cui et al. 2016). Incorporating the historical published literature poses a major challenge for this approach, however, because biologists have typically provided limited explicit guidance in their writings about what taxonomic concepts they have in mind for each instance of a name. Taken to the extreme, following Berendsohn's (1995) syntactic approach would require one to

[2] A taxonomic concept is a description of what a taxonomic name refers to as stated by a particular author in a particular publication. A taxonomic concept can be defined in terms of rules for appropriate use (an intensional definition), by a set of organisms included under the concept (an extensional definition), or by a mixture of these two approaches.

infer a taxononomic concept for each token use of each taxonomic name in the published literature.

This practical challenge for Berendsohn's (1995) approach reflects a deeper relationship between the syntactic and pragmatic approaches that arises from the different cognitive demands that humans and computers place on the functioning of taxonomic names. We show, for example, how biologists use names as metonyms to symbolize a variety of different objects, including type specimens, taxon concepts, and even historical progressions of taxon concepts. Such ambiguity of usage poses a major obstacle for making scientific discourse intelligible to machines because it exceeds, for example, even the expressive capacity of Berendsohn's (1995) extended syntax. For human readers with the appropriate expertise, however, ambiguity can increase efficiency by allowing taxonomists to reuse the same name for multiple purposes and in multiple contexts of scientific inquiry. There are thus important trade-offs involved in making scientific discourse intelligible to computers; weighing these trade-offs explicitly in relation to their underlying causes will better position the field to succeed than embracing the view that data aggregation works well enough without careful methodological guidance. We examine the merits of the syntactic and pragmatic strategies in terms of the trade-offs they make for how names can function as cognitive tools for coordinating research and facilitating communication across research groups and related organizations (Star and Griesemer 1989; Bowker 2000; Gerson 2008).

Big Data Trajectories and the Pitfalls of Aggregation

In general, "big data" evokes an optimism about progress: getting more data will make things better in the long run, even if it poses serious problems in the short run that require ongoing research and resources. Attempts to characterize big data in more detail have typically focused on producing theoretical definitions: claims that particular tools, methods, aims, or conditions are core to what makes big data. For example, big data is sometimes characterized in terms of the absolute number of data points, such as the billions of nucleotides produced by genome sequencing or the exabytes of data produced by telescopes in astronomy, along with the velocity (i.e., rate) and variety of data at issue (Laney 2001). Others focus on how scaling up data collection leads to demands for new technological infrastructure to handle all the data processing and storage (Hey et al. 2009), new data-driven analytical methods (Hey et al. 2009; Sepkoski 2012; Mayer-Schönberger and Cukier 2013; Meng 2014), acquiring truly complete datasets (Hutter and Moerman 2015), or transformation in social norms and career pathways (Strasser 2011; Lagoze 2014; Leonelli 2014).

Perhaps the most radical claim made on behalf of big data is that it renders previous scientific methods obsolete (Anderson 2008; Mayer-Schönberger and Cukier 2013). For instance, traditional concerns about random sampling and careful statistical analysis will be alleviated by having datasets that are complete, i.e., which include measurements made on every entity of interest in a domain. Common examples of "complete" datasets would include whole genome sequences or an exhaustive gene network that represents every causal interaction between pairs of genes in a cell.

We suggest it is more fruitful not to provide a single theoretical definition for several reasons. The most basic reason is that big data is still developing and expanding quite rapidly, so it is far from clear whether it has some determinate character at this point. Indeed, a second point is that big data is so popular in part because what it means is highly flexible, and it is a trendy topic for research funding or new business opportunities. Trying to give a precise definition for the phrase would direct our attention away from a key factor in its growth. Finally, attempting a single theoretical definition may obscure important complexities connecting the activity of generating large amounts of new data to changes in the epistemic tools and culture of a community. For example, it is unlikely that all disciplines involved in big data will sign up for the transformation of the scientific method advertised by Mayer-Schönberger and Cukier (2013), but Leonelli (2014, p. 9) suggests that "the real revolution seems more likely to centre on other areas of social life, particularly economics and politics, where the widespread use of patterns extracted from large datasets as evidence for decision-making is a relatively recent phenomenon." It is more important to analyze and explain these different outcomes of scaling up data than to focus only on the most radical cases.

As an alternative, we characterize big data using the idea of a "big data trajectory," which refers to the activities of a group of researchers who (1) set out to change the collective set of data available to address one or more shared problems of interest in such a way that (2) the researchers believe existing methods or resources available to the group are not adequate for the project and (3) they believe acquiring these methods or resources poses specific research problems separate from the original problems of shared interest.[3] Our emphasis on the trajectory of the researchers' activities is intended to highlight the sequence of decisions, obstacles, achievements, and outcomes they encounter

[3] For more on the concept of trajectory as a tool for comparative research in the social sciences, see Strauss (1993).

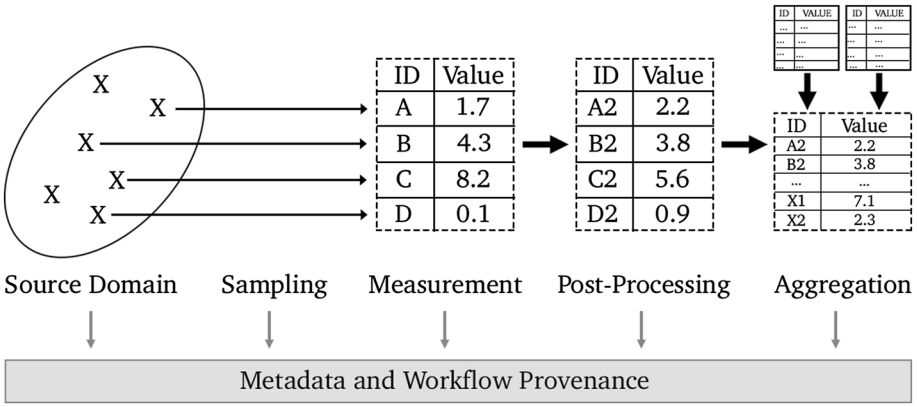

Fig. 1 Common structural elements in the process of data production. Each named element is an important source of methodological pitfalls that can lead to corruption or loss of value when scientists try to aggregate together datasets that were generated independently

during their project. In general, the idea serves as a nominal definition, i.e., it provides operational criteria for identifying cases of big data without attempting to select certain properties as most fundamental or explanatory. The point of such a definition is to constitute a set of cases for comparative analysis based on a shared standard that minimizes the influence of theoretical preconceptions on our understanding of the phenomena at issue (cf. Griesemer 2012).

It is a virtue of this approach that our definition of big data trajectories will include a number of cases that are not conventionally labeled as big data since these cases often provide useful insight into what features make other cases paradigmatic by contrast. Indeed, one can argue that the underlying phenomenon named by "big data" has been around for centuries in science (Ogilvie 2003; Müller-Wille and Charmantier 2012; Aronova et al. 2017). Although the absolute scale of information available today is unprecedented, the experience of a large relative increase in information is not new. For example, the large influx of new specimens that Europeans collected in the 1700s from around the world, many of which did not comfortably fit in existing taxonomies, drove Carl Linnaeus to formalize new taxonomic and nomenclatural practices in order to cope (Ogilvie 2003; Dietz 2012; Müller-Wille and Charmantier 2012; Charmantier and Müller-Wille 2014).

Given a set of big data trajectories, the goal is to investigate similarities and differences among the cases in order to identify factors that influence how individual trajectories or classes of trajectories develop over time. In this article, we focus on the consequences of theory-dependence for the task of data aggregation. In general, big data trajectories involve aggregating data, in the sense of bringing together pieces of data of a relevantly similar type into one location, e.g., a single database. The promise of big data that more is better is not uniformly accessible: scientists have found data aggregation to be incredibly difficult and time-consuming in fields where data are regularly produced across multiple independent organizational contexts (Edwards et al. 2011; Millerand et al. 2013; Leonelli 2014). The production of metadata—higher-order data about what is contained in each dataset—is a complex process involving negotiation and cross-disciplinary expertise, but without metadata a collection of data items remains just a sequence of strings and numbers in a text file or table somewhere (Edwards et al. 2011; Millerand et al. 2013). It turns out that the value of achieving big data does depend on how one gets there (boyd and Crawford 2012; Lagoze 2014; Lazer et al. 2014; Leonelli 2014). "The data that are most successfully assembled into big collections [in biology] are genomic data, such as genome sequences and microarrays, which are produced through highly standardised technologies and are therefore easier to format for travel" (Leonelli 2014, p. 5).

Figure 1 illustrates several key structural elements in data production processes that are loci for methodological pitfalls when it comes time to aggregate datasets. The post-processing stage reflects the need to estimate likely sources of error or bias in the initial values reported by observing researchers or measurement instruments (e.g., Hoeppe 2014). It may also be necessary to simplify a large database down to manageable proportions (Suciu 2013; Meng 2014). In this way, post-processing for some local use can eliminate signals or introduce biases that are relevant to answering other questions. Similarly, researchers or instruments producing the same sort of data may differ in the profile of errors they make, return results at different levels of precision, or even use incommensurable definitions of the "same" variable, such as spatial location (Shavit and Griesemer 2009, 2011).

Turning to sampling, there is no guarantee that a number of locally representative samplings will add up to a globally representative sample: datasets may partially overlap in coverage or have systemic gaps that were not relevant for their narrower uses. Finally, if users are missing metadata about any one of these aspects, this could render the aggregate dataset useless. Among other things, metadata provides a way of tracking the provenance of data back

through their production processes and of validating analytical assumptions about their properties.

One way to come to grips with the limitations of big data is therefore to investigate how aggregating datasets can fail to significantly increase their value for some purpose compared to leveraging each dataset on its own. Thinking in terms of big data trajectories allows us to encompass the diversity of circumstances, aims, and means within the movement while tracking how this diversity influences the benefits that communities pursuing big data end up realizing. In the next section, we use taxonomic name-based data in systematic biology to illustrate how aggregation can fail when the domain assigned to a single data point (a vouchered specimen) is a matter of theoretical inference in its own right.

Theory-Dependence in Biodiversity Data

Organisms preserved in museums or other collections are not much use to science unless they are identified as belonging to some taxon, usually at the species rank. However, definitions of species-level boundaries are scientific hypotheses,[4] so taxonomists will often reasonably disagree on which specimens should be identified to which taxonomic names. This means that aggregation necessarily involves more than placing an entry for each specimen record into a single database, since users typically request data by taxon name (or taxonomic concept label when available), or they request a list of specimens or taxa occurring within a geographic region under a higher-level name. Aggregation is thus theory-dependent in the sense that the correct aggregation of specimens under a name will typically require explicit tracking of how specimens have been identified according to various hypotheses about natural entities. More ambitiously, a database could track the correct assignments of specimens for each hypothesis and offer the user the ability to compare them.

In this section, we illustrate the importance of tracking taxonomic concepts to aggregating biodiversity data in order to motivate the theoretical importance of trade-offs between precision and ambiguity in taxonomic nomenclature. Despite the development of substantial databases and tools implementing the explicit tracking of taxonomic concepts (Ciardelli et al. 2009; Lepage et al. 2014; Cui et al. 2016), some experts continue to treat the issue as negligible or technically infeasible (e.g., Patterson et al. 2016). Our discussion therefore aims to expand on recent arguments for the importance and tractability of incorporating taxonomic concepts into biodiversity informatics.

Imagine you are the researcher responsible for building an on-line database that aggregates together records for all known specimens of one presumed species of grass, known by the valid taxonomic name *Andropogon virginicus*, along with any associated data published about the genomes, ecology, and other properties of the members of this entity. The ultimate goal is for each unique material specimen to have a single corresponding record in the database that includes all relevant information available about the specimen, including its taxonomic identification, date of collection, measured phenotypic or genetic properties, and geographic origin. For plants, a material specimen will often mean a pressed and dried cutting of one or more stems and leaves from an individual in the wild.[5]

The naïve solution would be to join together all records that link the name "*A. virginicus*"—coined by Linnaeus in 1753—to a specimen or dataset. Unfortunately, you quickly discover that, even after restricting the aggregation to (1) just in the period of 1889–2015 and (2) the region of the Mid-Atlantic and Southern United States, there have been 11 floristic and revisionary treatments that conflict over which organisms should be identified as *A. virginicus* or categorized as closely related under another name (Franz et al. 2016a).[6] Consequently, specimens used to describe the geographical range of *A. virginicus* as recognized in some hypothetical ecological study published in 1970 may no longer carry that name validly according to the most current classification today. Moreover, suppose that an influential flora for the region was published in 1968, yet the authors of the hypothetical study do not clearly indicate whether they are following this treatment or an incongruent earlier one (cf. Rosenberg 2014).

Certainly one cannot infer the relevant species membership of a specimen from the text string "*A. virginicus*" alone, because this string remains constant across all taxonomic revisions. We can momentarily set aside the additional complications of synonyms and spelling errors (Remsen 2016), because these do not alter the fundamental theory-dependency of name strings as reliable aggregators (Franz et al. 2016a). Perhaps some authors are careful enough to state which classification they are

[4] Note that these hypotheses are about the nature of individual species as entities, not the nature of biological species in general, which has been another source of ongoing debate among biologists and philosophers.

[5] More generally, we also include vouchered occurrence records, such as image-vouchered observations or tissue samples not linked to physical specimen depositions, under our use of the term "specimen data."

[6] While some of these treatments may provide nomenclatural synonymy information intended to resolve such conflicts, this information can nonetheless still be incomplete, incorrect, or out of date.

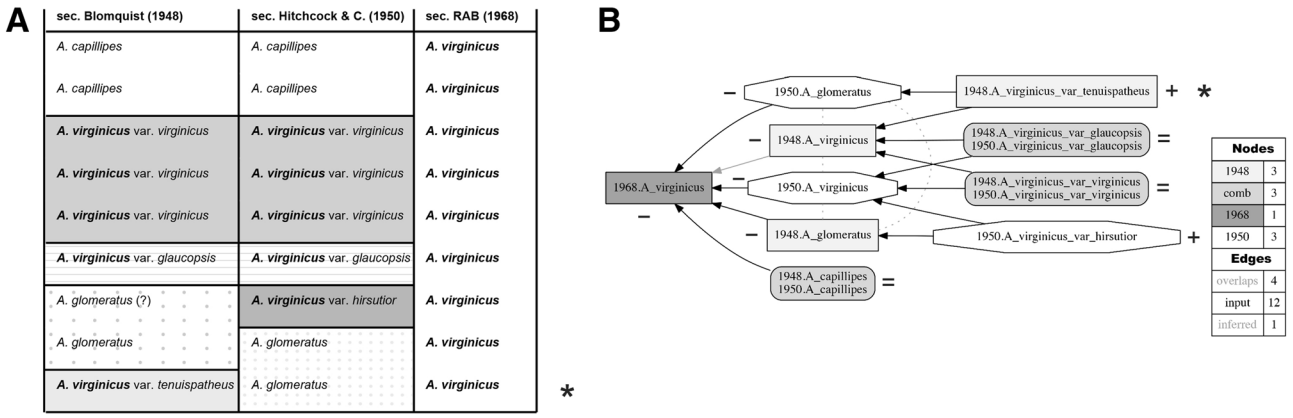

Fig. 2 Representations of the semantic relations between three classifications—Blomquist (1948), Hitchcock and Chase (1950), and Radford, Ahles, and Bell (1968)—utilizing the name *A. virginicus* (adopted from Franz et al. 2016a). **a** Tabular alignment of lowest-level taxonomic concepts; each horizontal row corresponds to a congruent lineage of taxonomic concepts labeled with different but (locally) valid names. **b** Directed acyclical graph alignment visualization ("meaning transition map") of the three classifications, as logically inferred with the Euler/X multi-taxonomy alignment toolkit (Franz et al. 2015, 2016a, b). Taxonomic concept labels are abbreviated. *Arrows* symbolize inverse proper inclusion (<), whereas *dashed lines* indicate overlap (><). Regions with *rounded corners* symbolize congruent concept regions; whereas other regions with various (*gray* to *white*) shadings and either *rectangular or octagonal* shapes represent concepts unique to the respective input taxonomy (i.e., either 1948, 1950, or 1968). Additional annotations: = congruent concepts, identical names; +: unique concept, unique name; − unique concept, non-unique name; * (**a** and **b**): low-level concept lineage congruent with *A. virginicus* var. *tenuispatheus* sec. Blomquist (1948), as singled out in the main text

using elsewhere in the publication, but this is more often the exception than the rule (Rosenberg 2014). Moreover, it turns out that many physical specimens were first identified and stored under prior classification schemes, and have not had their identifying name data updated since then (cf. Smith et al. 2016). So it looks as if building the database will routinely require you to sort through these ambiguities "by hand" for each data point used in thousands of papers and for hundreds of vouchered collections across institutions (Goodwin et al. 2015).

Now recall that all this effort is devoted to aggregating specimen data for a single species-level name, and that the ultimate goal is to build this sort of database for all such names, i.e., for all reliably inferred and named sections of the tree of life (though see O'Malley 2013).[7] Unfortunately, the reconciliation challenges posed by *A. virginicus* are not unusual across taxonomy (Franz et al. 2008). For instance, Geoffroy and Berendsohn (2003) found that 44.5% of the 1548 taxonomic names listed as valid in a reference list of German mosses published in 2000 showed instability in their taxonomic circumscription over a 73-year time period. Furthermore, the usages of many names may appear stable only because they have not received careful revision since their original definition (Remsen 2016).

How does this matter for data aggregation? If we count each specimen as a single datum, then what that datum is about is often a matter of theoretical dispute. In other words, taxonomic classifications reflect empirical hypotheses about which living things in the world belong to which evolutionarily coherent taxa, such that disputes about the nature of taxa are in part disagreements about what names the specimens in some collection are properly associated with. The very assignment of data to a source domain is therefore a matter of theory-dependent inference.

In consequence, attempting to aggregate specimen collections labeled under conflicting classifications can lead to pervasive errors in how data get assigned to sources. As one example of how errors can arise, consider a situation where two classifications agree on the boundaries of a group but disagree about how to partition it into species-level subgroups "A" and "B." Now consider a specimen that classification 1 assigns to the species name A, but which under classification 2 would be identified to the species name B. If we aggregate specimen collections from the two collections based on the identifiers A and B, then neither set of specimens in the aggregate dataset will have a determinate taxonomic meaning.

To properly address the integration challenge, we would have to implement a solution similar to that shown in Fig. 2 (cf. Franz 2016a). The names *A. virginicus* and *Andropogon glomeratus*, as used in the respective classifications of

[7] Our argument in this article does not presuppose that named taxa are monophyletic groups, but we will set aside the debate over how to define biological species as a further complication that only magnifies the difficulty of meaningful classification of specimen data.

1948 and 1950, have taxonomically intersecting meanings of the sort ascribed to A and B above. The 1968 classification in this example is less fine-grained than either of its predecessors. This could, for instance, mean that a specimen identified to the species-level name *A. virginicus* in 1968 would be correctly assigned to the name *A. glomeratus* in 1950, and to the variety-level name *A. virginicus* var. *tenuispatheus* in 1948, even though the 1948 classification treats *A. glomeratus* as a valid species-level name whose referent excludes that of *A. virginicus*.

We briefly point out that biologists may propagate such conflicting taxonomic name usages even when they all agree on the ostensive definitions of the names under consideration based on their nomenclatural type specimens (Witteveen 2015a, b).[8] In other words, biologists may all agree that the selection of a type specimen for (e.g.) a species-level name establishes a determinate fact of the matter about what the name designates; however, they may still disagree about what members *in addition to* the type specimen are included in the referent. In this regard, the intensional circumscription (e.g., feature-based, differential diagnosis) that taxonomists offer along with the type specimen to delimit the name's non-type reference serves as a scientific hypothesis about the taxon's evolutionary identity. Operating under the premise that there is such an identity in itself is not sufficient to tell us which currently advocated hypothesis (or hypotheses) among several competing options will be held as most valid in future taxonomic treatments.

Syntactic Solution: Explicit Tracking of Taxonomic Concepts

There are currently two main pathways available for resolving the source ambiguity for specimen data: syntactic and pragmatic (Remsen 2016). The former extends the internal grammar of taxonomic nomenclature in order to accommodate taxonomic concepts, whereas the latter leverages contextual information surrounding the usages of names in order to achieve more accurate resolutions of meaning. In order to understand the potential advantages and challenges of each strategy, we situate the problem in terms of tensions between the distinctive cognitive abilities of humans and computers. A syntactic solution would have considerable value for improving precision in future publications, but would fail to address the problem for legacy literature. As a consequence, there will be an ongoing need to develop new computational methods for incorporating pragmatic context into the semantic disambiguation of taxonomic names. This challenge offers an exciting and novel site for the intersection of informatics research in systematic biology and practice-oriented research in science studies.

A taxonomic concept reflects an empirical hypothesis on the part of the author about which organisms are in the taxon designated by the name. Taxonomic concepts are generally stated as a combination of intensional circumscriptions (i.e., descriptions) of the taxon's properties and specifications of individual specimens or lower-level concepts that match these properties. The idea to augment identifier resolution in biodiversity databases was formalized by Berendsohn (1995), who proposed that labels for taxonomic concepts should utilize the syntax, [taxonomic name] [name author and citation] *sec.* [source], where the "sec." stands for *secundum* (according to). An example of this syntactic convention would be "*A. virginicus* Linnaeus 1753: 1046s. Blomquist (1948)," which can thereby be differentiated from another taxonomic concept label "*A. virginicus* Linnaeus 1753: 1046s. Hitchcock and Chase (1950)."[9]

Several taxonomic publications and software projects have adopted Berendsohn's (1995) expanded syntax and implemented it to identify and track taxonomic concepts (e.g., Koperski et al. 2000; Pullan et al. 2000, 2005; Franz et al. 2015, 2016a, b; Jansen and Franz 2015; Cui et al. 2016). Avibase (Lepage et al. 2014) is one of the most thorough and high-impact realizations, storing unique labels for 844,000 species-level and 705,000 subspecies-level taxonomic concepts spanning across 151 checklists of birds published over 125 years. The database also tracks the extensional relationships between taxonomic concepts (Franz and Peet 2009), such as inclusion, inverse inclusion, equivalence, overlap, and exclusion. This added information facilitates the recognition of 38,755 taxonomically distinct concept "clusters"—i.e., uses of taxonomic names that correspond to the same circumscription of a taxon. An analysis of the Avibase data furthermore shows that only 11 of 19,260 (~1 in 1750) taxonomic name:concept combinations are both syntactically and semantically unique, i.e., these combinations have a symbol:reference cardinality of 1:1 across the entire Avibase taxonomic information environment. In all other cases, either one taxonomic name string has multiple incongruent meanings or several different name strings have one congruent meaning. In either case, this means that additional, human-facilitated contextual

[8] Interestingly, the type method is not mandatory above the family level where the codes of nomenclature have no regulatory power (cf. Franz and Thau 2010).

[9] Figure 2 illustrates the relationship between these two taxon concepts.

framing is needed to achieve precise name:concept assignments.

Returning to the theme of big data trajectories, we suggest that several special conditions may have catalyzed the (historically) early shift towards more precise syntactic design for taxonomic information in Avibase. For instance, Avibase focuses primarily on representing classifications below the family level, where precise reference is often regarded as critical (Peterson and Navarro-Sigüenza 1999). The database must accommodate one or more new taxonomies each year, published as updated checklists by different global or regional authorities such as the American Ornithologists' Union or the International Ornithological Committee. Invariably, there are significant changes in name usage across checklists. To reduce the complexity of the task, each represented checklist is a coherent, time-stamped perspective with a simple and consistent data format. Avibase tracks neither feature-based circumscriptions nor individual specimen data. There is immense scientific and public use of these evolving taxonomies, whose syntactic and semantic complexities rapidly exceed an individual human's cognitive abilities for name:concept reconciliation. It is as if for once humans were "put in the shoes of computers," and realized that the desired multi-taxonomy aggregation services could only be obtained through more context-aware name usage identifiers.

Pragmatic Solution: Analyzing Ambiguity in Context

Adding a standardized suffix to taxonomic names disambiguates taxonomic concepts by extending their internal syntax in order to replace the contextual information typically used by human readers (Rosenberg 2014). However, even if this new convention could be universally implemented, it would not solve the problem of disambiguating all name usages and taxonomic concepts published in the past. In order to incorporate the massive legacy data about specimens from the historical literature, it seems necessary to introduce some sort of natural language processing approach—i.e., use computational algorithms to make inferences about intended meanings based on textual information surrounding particular uses of names (Cui et al. 2016; Gandy et al. 2016). Thus far in the article, we have considered the semantics of taxonomic names largely in the already tame context of aggregating pre-existing data records. When we turn to look at taxonomic names "in the wild," it turns out that even Berendsohn's (1995) "sec." convention is inadequate to disambiguate the full range of entities a taxonomic name can symbolize. In particular, it is quite common for biologists to use a name to ascribe properties like stability or instability to a trajectory of taxonomic concepts over time. As a result, the pragmatic context of a taxonomic name can demand one of several possible models for semantic disambiguation.

Consider a few possible statements taxonomists might say or write to their colleagues, each of which are perfectly adequate instances of English in an appropriate context.[10] (1) "*A. virginicus* is in cabinet 5, shelf 3." (2) "*A. virginicus* is specimen LINN 1211.12." (3) "*A. virginicus* occurs only in the eastern United States." (4) "Radford, Ahles, and Bell's (1986) definition of *A. virginicus* was based on the biological species concept." (5) "Although initially poorly understood, *A. virginicus* has settled into a stable definition as a result of extensive study over the past five decades." In each case, the same name string "*A. virginicus*" symbolizes the following referents: (1) a collection of one or more specimens, (2) the designated type specimen for the species name (epithet), (3) the taxon itself (as recognized at the time), (4) a particular taxonomic concept, and (5) a temporal sequence of multiple taxonomic concepts.

While sentences (1) and (2) are more colloquial and unlikely to show up in the scientific literature, [11] instances of sentences like (3)–(5) are highly common in the published systematic literature, especially when legacy and new, taxonomically referenced information is brought together (e.g., Gratton et al. 2016).

In linguistic terms, these cases illustrate how biologists use the taxonomic name *A. virginicus* as a metonym to stand for a variety of related entities. Metonymy is "a figure of speech consisting of the use of the name of one thing for that of another of which it is an attribute or with which it is associated (as 'crown' in 'lands belonging to the crown')" (Merriam-Webster 2016). Most importantly for our purposes, use (5) for taxonomic names shows how natural language "in the wild" invokes meanings that go beyond the immediate expressive capacity of the extended taxonomic concept label syntax, i.e., [name] sec. [source].

Cognitive Pragmatics for Taxonomic Names

In the face of ambiguity about the identification of data, a common response among biodiversity informaticians is to push for a single, universal system that applies unique names to each lowest-level unit of data (cf. Godfray 2002). The issue of theory-dependence here operates at one level of composition higher: given a system of unique identifiers for each physical specimen, there is still significant

[10] These statements are hypothetical examples and should not be taken as necessarily true.

[11] Although see Figures 3C and 3D in Remsen (2016) for a visual analog to sentence (2).

ambiguity about how to group those specimen records into biologically relevant units. One can imagine reiterating the same response, then, at the level of taxonomic nomenclature: implement a new or revised set of nomenclatural rules so that names encode all the information needed to disambiguate which taxonomic concept the author is using. This strategy reflects a broader stance that ambiguity is generally a hindrance to science and that a universal system of standards and practices that minimizes ambiguity is both possible and preferable. This stance takes for granted that the validity of the system of standards and practices is largely independent of the ongoing results of the research it facilitates. Theory-dependence in data threatens this stability insofar as it introduces conflict and uncertainty into the basic properties of the data, including their identity and semantics. This challenges a familiar conception of science, and of big data trajectories in particular, as making consistent progress through the accumulation of ever more bits of data (cf. Kuhn 1996[1962]).

A positive role for ambiguity in this regard is to enable terms to mediate between conflicting positions by providing a common ground that leaves undecided key issues for debate or negotiation. Indeed, this is one virtue of taxonomic names such as *A. virginicus*, which do not specify the associated taxonomic concept or species category being invoked. Anchoring a name ostensively using a type specimen provides a common ground for scientific debate over the nature of the taxon without building a resolution to that debate into the name itself. A system for identifying taxonomic concepts that replaced Linnaeus's binomial system with a 128-bit sequence of zeroes and ones would achieve a new level of precision at the cost of our ability as humans to use taxonomic names to mediate disputes over which taxonomic concept is correct.[12] We also saw how multiple uses of the name as a metonym reflect the expressive power enabled by ambiguity that would be eliminated by embracing a more restrictive syntax.

In this vein, we argue that any given nomenclatural convention carries one or more cognitive trade-offs for its users. From Linnaeus onward, the efficiency, stability, and precision of communication have been leading motivations for the design and revision of practices and rules for biological nomenclature (Stevens 2002; Ogilvie 2003; Dayrat 2010; Dietz 2012; Müller-Wille and Charmantier 2012; Witteveen 2015b). For example, one reason Linnaeus was motivated to formalize the binomial system was to escape a common expectation at the time that species-level names would include enough descriptive information to distinguish the purported species from all con-generic members (Cain 1958; Stearn 1959; Jansonius 1981). As European explorers began to collect ever more specimens from diverse sections of the tree of life, this expectation forced taxonomists to keep adding more descriptive content to species-level names, thereby allowing them to adequately distinguish their meanings from those of presumed close relatives. For example, between 1738 and 1753 Linnaeus was forced to expand the name string "*Plantago foliis ovatis glabris*" into "*Plantago folios ovaries glares, nudo scapo tereti, spica flosculis imbricatis*." Shortly thereafter, he simplified it to the more manageable "*Plantago major*" (Jansonius 1981).

Biologists' taxonomic naming practices are thus a particularly regimented example of what is arguably a very general trade-off between ambiguity and precision in human language (Zipf 1949; Atran 1998; Levinson 2000; Wilson and Sperber 2012). Many linguists and philosophers have assumed that the optimal design for a communication system is perfect semantic precision, such that each word corresponds to only one possible meaning (e.g., Chomsky 2002). However, when people share elements of their environmental context, this precision actually leads to inefficient redundancies. According to Piantadosi et al. (2012, p. 281; emphasis in original), "ambiguity is in fact a *desirable* property of communication systems, precisely because it allows for a communication system which is 'short and simple.'" These authors argue for "two beneficial properties of ambiguity: first, where context is informative about meaning, unambiguous language is partly redundant with the context and therefore inefficient; and second, ambiguity allows the re-use of words and sounds which are more easily produced or understood" (Piantadosi et al. 2012, p. 281).

Although Piantadosi et al. (2012) examine the value of ambiguity in a static or equilibrium context across whole languages, their reasoning should also have implications for the dynamics of linguistic change. In this vein, the attempt to assemble big data in systematics is likely to destabilize taxonomic nomenclature by forcing new compromises between the divergent cognitive abilities of humans and computers to parse the context-specific meanings of taxonomic names. In other words, the drive to big data will force the issue of whether taxonomy is best designed for humans or for computers.

While expert taxonomists have successfully dealt with the semantic ambiguities of names for centuries, the strategies they have developed do not scale well across millions of biological entities compounded by hundreds of years of history. Indeed, an important part of taxonomists' expertise lies in having committed to memory the historical sequence of taxonomic revisions and nomenclatural relationships for some domain of interest, along with the idiosyncrasies

[12] Note that Berendsohn's extended syntax maintains coherence with existing practices in taxonomy by adding onto the binomial system rather than replacing it wholesale.

of collections housed at different institutions. Thus, when experts read a published article about some taxonomic group, they are typically able to use contextual clues from the text plus their extensive, fine-tuned background knowledge of the domain to correctly parse the authors' name usages and concepts of the corresponding taxa. Unfortunately, relying on context to resolve ambiguity poses a profound challenge for computers, which presently lack humans' cognitive skills for making complex inferences about the intended meanings of names.

In the opposite direction, computational logic is not immediately optimized for human communication and reasoning about taxonomic entities. Hypothetically speaking, if the objective is to logically represent a transition between taxonomic concepts over time, computers have no need to reuse any symbols for references authored at separate times. Moreover, the symbols would not need to be mono- or binomial, or of limited length and thereby easily pronounceable and memorable for humans. Rank endings would not have to be inherently embedded in the symbols, as this information could be allocated in an associated table. Instead it would be perfectly suitable in the "eyes" of computational representation to assign (e.g.) globally unique 128-bit strings for every name usage. Handling trillions of such unique symbols to account for taxonomic concepts in the systematic knowledge domain would be well within reach in terms of computational processing capacity.

Assigning unique identifiers to taxonomic concepts in this way may approximate optimality for computers, but it does so at the expense of rendering human communication about taxonomic names, referents, and provenance far more difficult. In particular, the positive scientific roles served by the ambiguity of taxonomic names would be lost or would have to be reconstructed using a different syntactic convention. Hence, we need to ask: is there an attainable and desirable middle-ground solution? Is the "best of both worlds"—i.e., taxonomic symbols that are maximally aligned with human cognitive capacities yet that are also precisely framed for logic-based representation and reasoning—an option? In order to answer this question, we need to make explicit the trade-offs between precision and ambiguity for systems used by both humans and computers to identify and classify data. Given the default tendency of many scientists and philosophers to prefer precision, this implies the need for new attention to the positive roles that ambiguity plays in scientific communication and inquiry.

Conclusion

Criticism of the big data movement has largely focused on showing that more is not necessarily better, in the sense that data do not always maintain their value when taken out of their immediate context of use and aggregated together. Nonetheless, there is a value and even imperative to pursuing big data in many fields, such as for biodiversity studies. The next question, then, is how to incorporate an awareness of aggregation pitfalls into the design of the infrastructure and institutions supporting a group's big data trajectory. One common strategy aims to minimize aggregation pitfalls by increasing the precision of our conventions for identifying and classifying data. The theory-dependence of specimen data is not simply an obstacle for aggregation, however, because it also reflects the status of taxonomic concepts as scientific hypotheses subject to continuing research and debate. We argued for the positive roles that ambiguity can play in this regard as an efficient way of symbolizing multiple aspects of taxa and mediating between different conceptions of their nature. We also argued that trade-offs between precision and ambiguity are a general feature of taxonomic nomenclature given the multiple functions taxonomic names serve in biological research. Ambiguity is therefore not purely an obstacle to scientific progress, but has positive contributions to make that should be balanced against precision in the design of data infrastructure.

The existence of trade-offs between precision and ambiguity is connected to fundamental questions about the nature of scientific change in the history, philosophy, and social studies of science. Behind many of the promises made for big data is the assumption that collecting more data can lead to major changes in science without challenging the meaning or value of that same data. In other words, very few scientists and funding agencies engaged in building huge new databases expect to simply throw them away once those databases have generated their important new results.

This reflects a familiar understanding of scientific progress as a steady accumulation of knowledge: "If science is the constellation of facts, theories, and methods collected in current texts, then scientists are the men who, successfully or not, have striven to contribute one or another element to that particular constellation. Scientific development becomes the piecemeal process by which these items have been added, singly and in combination, to the ever growing stockpile that constitutes scientific technique and knowledge" (Kuhn 1996[1962]), p. 2). What is new here in the context of big data is that informatics has brought a conceptual precision to the handling and use of data that has historically been applied mainly to the analysis of scientific theories. What Kuhn showed us, however, is that the development of science over time includes discontinuities, contradictions, and shifts in perspective that cannot be dismissed as simply the ignorance, errors, or superstition of

the past. Even the most logically precise understanding afforded by our current knowledge, then, is not a sufficient guide to the meaning and value of past or future data.

A higher synthesis of the virtues of precision and ambiguity would go beyond the demonstration of trade-offs to characterize their systematic interdependence within the circumstances of a big data trajectory. For example, making explicit taxonomic concept labels a universal standard could serve as the basis for a fruitful synergy between the syntactic and pragmatic approaches we described. Since being able to make assertions about how taxonomic concepts have changed over time is essential for biologists to describe advances in empirical support for emerging classificatory patterns across multiple revisions, it would be useful to express these changes in formal terms. Taxonomic concept labels make this possible by allowing one to reconstruct claims about how scientific knowledge about a taxonomic "region" has changed in terms of logical relations between taxonomic concepts (Fig. 2), and leverage these assertions for semi-automated, scalable logic applications (Cui et al. 2016; Franz et al. 2016b, c).

Similarly, when scientists regiment natural language by adopting new syntactic or semantic conventions, it has downstream effects on the pragmatics of future discourse. For example, adopting Berendsohn's (1995) "sec." extension to traditional nomenclature would force the authors of a new study to routinely decide, with every name usage, whether they are (1) authoring a concept with relevantly new semantics, (2) referring to a name usage that they concur with (perhaps inaccurately) but that is "by someone else," or (3) not really referring to any particular usage (explicitly wanting to commit to non-precision). Only (1) requires coining a new identifier, but deciding whether that is merited or not is possibly a new burden, and a subject not yet well explored. In other words, the opportunity to specify name usages more precisely translates into a heightened responsibility, or demand, to make one's speaker role more conscious and explicit. As these examples suggest, interesting and important challenges for designing the data infrastructures of science arise where the virtues of precision and ambiguity interact. Pursuing a deeper understanding of the trade-offs and synthesis of these two aspects of scientific language therefore offers a next step for the productive critique and design of big data systems.

Acknowledgements The authors are grateful to Hong Cui, Bertram Ludäscher, and Jonathan Rees for helpful feedback on this subject. Support of the authors' research through the National Science Foundation is kindly acknowledged (NMF: DEB–1155984, DBI–1342595; BS: SES–1153114).

References

Anderson C (2008) The end of theory: the data deluge makes the scientific method obsolete. Wired, 23 June. http://www.wired.com/2008/06/pb-theory/

Aronova E, von Oertzen C, Sepkoski D (eds) (2017) Data histories, vol 32. Osiris, New York (in press)

Atran S (1998) Folk biology and the anthropology of science: cognitive universals and culturalparticulars. Behav Brain Sci 21:547–569

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

Berman JJ (2013) Principles of big data. Elsevier, Waltham

Blomquist HL (1948) The grasses of North Carolina. Duke University Press, Durham

Bowker GC (2000) Biodiversity datadiversity. Soc Stud Sci 30:643–683

boyd D, Crawford K (2012) Critical questions for big data. Inform Commun Soc 15:662–679

Cain AJ (1958) Logic and memory in Linnaeus' system of taxonomy. Proc Linn Soc Lond 169:144–163

Charmantier I, Müller-Wille S (2014) Carl Linnaeus's botanical paper slips (1767–1773). Intellect Hist Rev 24:215–238

Chomsky N (2002) An interview on minimalism. In: Belletti A, Rizzi L (eds) On nature and language. Cambridge University Press, Cambridge, pp 92–161

Ciardelli P, Kelbert P, Kohbecker A et al (2009) The EDIT platform for cybertaxonomy and the taxonomic workflow: selected components. Lect Notes Inform 154:625–38

Cui H, Xu D, Chong SS et al (2016) Introducing explorer of taxon concepts with a case study on spider measurement matrix building. BMC Bioinform 17(1):471

Dayrat B (2010) Celebrating 250 dynamic years of nomenclatural debates. In: Polaszek A (ed) *Systema Naturae* 250: The Linnean ark. CRC Press, Boca Raton, pp 186–239

Dietz B (2012) Contribution and co-production: the collaborative culture of Linnaean botany. Ann Sci 69:551–569

Edwards PN, Mayernik MS, Batcheller AL et al (2011) Science friction: data, metadata, and collaboration. Soc Stud Sci 41:667–690

Franz NM, Peet RK (2009) Towards a language for mapping relationships among taxonomic concepts. Syst Biodivers 7:5–20

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

Franz NM, Peet RK, Weakley AS (2008) On the use of taxonomic concepts in support of biodiversity research and taxonomy. In: Wheeler QD (ed) The new taxonomy. CRC Press, Boca Raton, pp 63–86

Franz NM, Chen M, Yu S et al (2015) Reasoning over taxonomic change: exploring alignments for the *Perelleschus* use case. PLoS ONE 10(2):e0118247

Franz NM, Chen M, Kianmajd P et al (2016a) Names are not good enough: reasoning over taxonomic change in the *Andropogon* complex. Semant Web 7:645–667

Franz NM, Pier NM, Reeder DM et al (2016b) Two influential primate classifications logically aligned. Syst Biol 65:561–582

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

Gandy L, Gumm J, Fertig B et al (2016) Synthesizer: expediting synthesis studies from context-free data with natural language processing. bioRXiv. doi:10.1101/053629

Geoffroy M, Berendsohn WG (2003) The concept problem in taxonomy: importance, components, approaches. Schrift Vegetationsk 39: 5–14

Gerson EM (2008) Reach, bracket, and the limits of rationalized coordination: some challenges for CSCW. In: Ackerman MS, Halverson CA, Erickson T, Kellogg WA (eds) Resources, co-evolution and artifacts: theory in CSCW (computer supported cooperative work). Springer, London, pp 193–220

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

Goodwin ZA, Harris DJ, Filer D et al (2015) Widespread mistaken identity in tropical plant collections. Curr Biol 25:R1066–R1067

Gratton P, Trucchi E, Trasatti A et al (2016) Testing classical species properties with contemporary data: how "bad species" in the brassy ringlets (*Erebia tyndarus* complex, Lepidoptera) turned good. Syst Biol 65: 292–303

Griesemer JR (2012) Formalization and the meaning of 'theory' in the inexact biological sciences. Biol Theory 7(4):298–310

Hey T, Tansley S, Tolle K (eds) (2009) The fourth paradigm: data-intensive scientific discovery. Microsoft Research, Redmond

Hinchcliff CE, Smith SA, Allman JF et al (2015) Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc Natl Acad Sci USA 112:12764–12769

Hitchcock AS, Chase A (1950) Manual of the grasses of the United States, 2nd edn. United States Department of Agriculture Miscellaneous Publication No. 200. US Department of Agriculture, Washington, DC

Hoeppe G (2014) Working data together: the accountability and reflexivity of digital astronomical practice. Soc Stud Sci 44:243–270

Hutter H, Moerman D (2015) Big data in *Caenorhabditis elegans: quo vadis*? Mol Biol Cell 26:3909–3914

Jansen MA, Franz NM (2015) Phylogenetic revision of *Minyomerus* Horn, 1876s. Jansen & Franz, 2015 (Coleoptera, Curculionidae) using taxonomic concept annotations and alignments. ZooKeys 528:1–133

Jansonius J (1981) Linnaean nomenclature. Universal language of taxonomists. And the *Sporae Dispersae* (with a commentary on Hughes' proposal). Taxon 30:438–448

Koperski M, Sauer M, Braun W, Gradstein SR (2000) Referenzliste der Moose Deutschlands. Schrift Vegetationsk 34:1–519

Kuhn TS (1996[1962]) The structure of scientific revolutions. University of Chicago Press, Chicago

Lagoze C (2014) Big data, data integrity, and the fracturing of the control zone. Big Data Soc 1(2):1–11

Laney D (2001) 3D data management: controlling data volume, velocity, and variety. Application Delivery Strategies, META Group Inc, Atlanta

Lazer D, Kennedy R, King G, Vespignani A (2014) The parable of Google flu: traps in big data analysis. Science 343:1203–1205

Leonelli S (2014) What difference does quantity make? On the epistemology of big data in biology. Big Data Soc 1(1):1–11

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

Levinson SC (2000) Presumptive meanings: the theory of generalized conversational implicature. MIT Press, Cambridge

Mayer-Schönberger V, Cukier K (2013) Big data: a revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt, New York

Meng X-L (2014) A trio of inference problems that could win you a Nobel Prize in statistics (if you help fund it). In: Lin X, Genest C, Banks DL et al (eds) Past, present, and future of statistical science. CRC Press, Boca Raton, FL, pp 537–562

Merriam-Webster (2016) Metonymy. http://www.merriam-webster.com/dictionary/metonymy. Accessed 13 Jan 2017

Millerand F, Ribes D, Baker KS, Bowker GC (2013) Making an issue out of a standard: storytelling practices in a scientific community. Sci Technol Hum Values 38:7–43

Müller-Wille S, Charmantier I (2012) Natural history and information overload: the case of Linnaeus. Stud Hist Philos Biol Biomed Sci 43:4–15

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

Ogilvie BrW (2003) The many books of nature: renaissance naturalists and information overload. J Hist Ideas 64:29–40

Page RD (2016) Surfacing the deep data of taxonomy. Zookeys 550:247–260

Patterson D, Mozzherin D, Shorthouse D, Thessen A (2016) Challenges with using names to link digital biodiversity information. Biodivers Data J 4:e8080. doi:10.3897/BDJ.4.e8080.

Peterson AT, Navarro-Sigüenza AG (1999) Alternate species concepts as bases for determining priority conservation areas. Conserv Biol 13:427–431

Piantadosi ST, Tily H, Gibson E (2012) The communicative function of ambiguity in language. Cognition 122:280–291

Pullan MR, Watson MF, Kennedy JB et al (2000) The Prometheus taxonomic model: a practical approach to representing multiple classifications. Taxon 49:55–75

Pullan MR, Armstrong KE, Paterson T et al (2005) The Prometheus description model: an examination of the taxonomic description-building process and its representation. Taxon 54:751–765

Radford AE, Ahles HE, Bell CR (1968) Manual of the vascular flora of the Carolinas. University of North Carolina Press, Chapel Hill

Remsen D (2016) The use and limits of scientific names in biological informatics. ZooKeys 550:207–223

Rogers N (2016) Museum drawers go digital. Science 352:762–765

Rosenberg MS (2014) Contextual cross-referencing of species names for fiddler crabs (genus *Uca*): an experiment in cyber-taxonomy. PLoS ONE 9(7):e101704

Sepkoski D (2012) Rereading the fossil record: the growth of paleobiology as an evolutionary discipline. University of Chicago Press, Chicago

Shavit A, Griesemer JR (2009) There and back again, or the problem of locality in biodiversity surveys. Philos Sci 76:273–294

Shavit A, Griesemer JR (2011) Transforming objects into data: how minute technicalities of recording 'species location' entrench a basic challenge for biodiversity. In: Carrier M, Nordmann A (eds) Science in the context of application. Boston Studies in the Philosophy of Science, vol 274. Springer Science + Business Media, Netherlands, pp 169–193

Smith BE, Johnston MK, 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

Star SL, Griesemer JR (1989) Institutional ecology, 'translations' and boundary objects: amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. Soc Stud Sci 19:387–420

Stearn WT (1959) The background of Linnaeus's contributions to the nomenclature and methods of systematic biology. Syst Zool 8:4–22

Stevens PF (2002) Why do we name organisms? Some reminders from the past. Taxon 51:11–26

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

Strauss AL (1993) Continual permutations of action. de Gruyter, New York

Suciu D (2013) Big data begets big database theory. In: Gottlob G, Grasso G, Olteanu D, Schallhart C (eds) Proceedings of the 29th British National Conference on Databases, BNCOD 2013,

Oxford, UK, July 8–10, 2013. Spring, Berlin. Lect Notes Comput Sci 7968:pp 1–5
Wilson D, Sperber D (2012) Meaning and relevance. Cambridge University Press, New York, NY
Witteveen J (2015a) Naming and contingency: the type method of biological taxonomy. Biol Philos 30:569–586
Witteveen J (2015b) Suppressing synonymy with a homonym: the emergence of the nomenclatural type concept in nineteenth century natural history. J Hist Biol 49:135–189
Zipf G (1949) Human behavior and the principle of least effort. Addison-Wesley, New York