TAXONOMIC CHECKLISTS AS BIODIVERSITY DATA: HOW SERIES OF CHECKLISTS CAN PROVIDE INFORMATION ON SYNONYMY, CIRCUMSCRIPTION CHANGE AND TAXONOMIC DISCOVERY

By

GAURAV GIRISH VAIDYA

B.Sc., National University of Singapore, 2006

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirement for the degree of
Doctor of Philosophy
Department of Ecology and Evolutionary Biology
2017

This thesis entitled:
Taxonomic checklists as biodiversity data: how series of checklists can provide information on synonymy, circumscription change and taxonomic discovery
written by Gaurav Girish Vaidya
has been approved for the Department of Ecology and Evolutionary Biology

_____

Robert Guralnick

_____

J. Patrick Kociolek

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Vaidya, Gaurav Girish (Ph.D., Ecology and Evolutionary Biology)

Taxonomic checklists as biodiversity data: how series of checklists can provide information on synonymy, circumscription change and taxonomic discovery

Thesis directed by Professor Robert Guralnick

Taxonomic checklists are a fundamental and widely-used product of taxonomy, providing a list of recognized taxa within a taxonomic group in a particular geographical area. Series of taxonomic checklists provide snapshots of recognized taxa over a period of time. Identifying and classifying the changes between these checklists can provide information on rates of name, synonym and circumscription change and can improve aggregation of datasets reconciled to different checklists.

To demonstrate this, I used a series of North American bird checklists to test hypotheses about drivers of splitting rates in North America birds. In particular, I asked if splitting was predominantly undoing previous lumping that happened during the heyday of the modern synthesis. I found that bird species have been split at an accelerating rate since the 1980s. While this was partially the result of previously lumped species being resplit, most splits were unrelated to previous lumps and thus represent new discoveries rather than simply the undoing of previous circumscription changes. I also used a series of North American freshwater algal checklists to measure stability over fifteen years, and found that 26% of species names were not shared or synonymized over this period. Rates of

synonymization, lumping or splitting of species remained flat, a marked difference from North American birds. Species that were split or lumped (7% of species considered) had significantly higher abundance than other species in the USGS NAWQA dataset, a biodiversity database that uses these checklists as an index. They were associated with 19% of associated observations, showing that a small number of recircumscribed species could significantly affect interpretation of biodiversity data.

To facilitate this research, I developed a software tool that could identify and annotate taxonomic changes among a series of checklists, and could use this information to aggregate biodiversity data, which will hopefully facilitate similar research in the future. My dissertation demonstrates the value of taxonomic checklists series to answer specific questions about the drivers of taxonomic change ranging from philosophical and technical changes to characteristics of species themselves such as their abundance.

ACKNOWLEDGMENTS

I would like to thank my parents, family and friends for their support and love over the long, long path to this dissertation. I would particularly like to thank my dad, mum, sister, Caitlin Kelly, Christine Avena, Brian Putnam, Kim Schoonover, Melinda Markin, Kevin Bracy Knight, Amber Churchill, Helen McCreery, Amanda Hund, Tim Szewczyk, Sierra Love Stowell, Julie Allen, Daisie Huang and Denise Tan. Aspects of my PhD relied on the assistance, mentorship, advice and on conversations with Hilmar Lapp, Nico Franz, Trish Rose-Sandler, William Ulate, Matt Yoder, Andrea Thomer, Dimitris Kontokostas, Dean Pentcheff, John Wieczorek and Walter Jetz. I would particularly like to thank Pat Kociolek, Erin Tripp, Andrew Martin, Andrew M. Johnson, Nico Cellinese and Hilmar Lapp for their patience, advice and support.

As a PhD student, I have had many, many opportunities to work on fantastic projects unrelated to my core dissertation topic. I would like to thank the Wikimedia Foundation, NESCent, the Biodiversity Heritage Library, the Google Summer of Code project, and MCN for these opportunities. I am grateful to Nolan Kane and John M. Basey for the opportunity to teach undergraduate labs under their supervision. I would like to especially mention the developers of Java's Stream API, without which this dissertation would have taken much longer.

For Chapter 2. The tempo and mode of the taxonomic correction process in North American Birds over the last 127 years**:** Victoria Tersigni helped collect species description dates for this paper. I would like to thank Carla Cicero, Nico

Franz and John Bates for their feedback and comments on previous drafts of this manuscript and for Maxwell Joseph's comments on the hierarchical model. My initial work on this project was funded by a graduate fellowship at the National Evolutionary Synthesis Center under the supervision of Hilmar Lapp.

Most importantly: this dissertation could not have been begun, let alone seen to completion, without the advice, encouragement, feedback, criticism, inspiration and support of my Ph.D. advisor, Robert Guralnick. Thank you so much for everything, Rob.

CONTENTS

TABLES

FIGURES

CHAPTER 1. INTRODUCTION

Few aspects of biology are as universal or as fundamental as the description and classification of biological taxa. Whether measuring traits in a population, quantifying the amount of unique biodiversity in an area, or collecting information on a set of related individuals, scientists often begin their study by identifying taxa of interest, looking up previously accumulated knowledge about those taxa, and determining how to distinguish those taxa from others. This process is possible because of the work of taxonomists, who produce definitions of taxonomic units and assign unique taxonomic names to them as identifiers. These identifiers are widely used to refer to biological entities by scientists, conservationists, and members of the public, and may be published in scientific journals, acts of legislation, textbooks, field guides and encyclopedias. Despite its use in many different environments by many different users, biological nomenclature remains a global, unified system that produces identifiers that are, by and large, used consistently. These names are created and maintained under the provisions of nomenclatural codes (Jach 2000; McNeill et al. 2012; Lapage et al. 1992), which are adopted and enforced by international bodies and mandated by scientific journals, establishing a globally accepted convention that almost all biologists adopt.

New taxonomic names are established through taxonomic descriptions: formal nomenclatural acts that provide evidence supporting a novel, unnamed taxon as well as a name to be used for that taxon in the future. Studies of taxonomic process usually focus on descriptions, whether studying the rate at which

descriptions of currently recognized species have accumulated (Costello, Wilson, and Houlding 2012), how the length and complexity of descriptions have changed over time (Sangster and Luksenburg 2015), or how often described species are judged to be synonyms of other names (Gaston and Mound 1993). These have been used to extrapolate the number of species remaining to be described (Costello et al. 2015), to measure how quickly description rates are rising (Joppa, Roberts, and Pimm 2011; Tancoigne and Dubois 2013), and to determine where, how and by whom these descriptions are being made (Tancoigne et al. 2011). These studies deepen our understanding of the 267 year history of modern taxonomy, dated from the publications of Linnaeus' *Species Plantarum* (Linnaeus 1753) and the 10th edition of *Systema Naturae* (Linnaeus 1758), and can provide insight into the future trajectory of taxonomic description.

However, descriptions are seldom used directly by scientists in identifying taxa or aggregating biodiversity data. Conceptions of what a taxon is and what evidence it should be based on have changed significantly over the last 250 years (Haffer 1992) – for example, Linnaeus' initial publications of binomial names predate the development of the theory of natural selection by over a hundred years (Darwin 1859). This is particularly important at the species level, where at least 24 species concepts have been proposed and debated, many of which use different criteria to determine how species should identified and described (De Queiroz 2007). Furthermore, all taxonomic descriptions are hypotheses that may be rejected or emended in light of changing conceptions and new evidence (Sluys 2013). If the goal

of taxonomy is to provide a complete, accurate, stable organization of global biodiversity into clearly defined taxonomic groups, then the point at which taxonomic work can be said to be complete is not once all taxa have been described, but once taxonomic hypotheses have been tested and synthesized into a single, consistent taxonomy that is no longer being debated by the community.

So how close are we to that goal, and how long might it take to get there? Broadly, answering this question requires an understanding of what taxonomic activity takes place after original description: what changes are proposed by taxonomists, how many of those changes become generally accepted, and how those changes are themselves changed in the future. If we had complete information on every aspect of this process, we could trace the path of every taxonomic opinion from one circumscription to another, quantify what this path looks like for any average taxon, and compare this process between different groups, different taxonomic philosophies or different techniques. These opinions are generally scattered throughout the taxonomic literature, and except where some of them are aggregated into databases (Alroy 2002), require considerable effort to synthesize from disparate taxonomic sources (Sangster 2009, 2014). There is no way to be sure that every taxonomic opinion has been collected, and identifying how each opinion related to every other requires considerable effort (see e.g. Lepage, Vaidya, and Guralnick 2014).

We can reframe this question more narrowly by focusing on currently recognized taxa. This sets aside historical patterns of taxonomic change, which may

be useful in understanding how taxonomy functions as a process, in order to focus on how specific changes in taxonomic practice – such as a shift from the biological species concept to a phylogenetic species concept (Agapow et al. 2004) – are affecting the recognition of taxa today. The pace at which these changes are taking place is critical to estimating what their future effect might be: a high rate of change in the past might indicate that current taxa are approaching stable definitions, for example, while ongoing change might suggest that current taxa are far from stabilizing. This question has immediate practical significance in understanding how stable our current taxonomic view is, how specific changes in taxonomic practice might make circumscriptions more or less likely to change, and may still provide insights into how taxonomic redescription generally works.

Taxonomic names are often compiled into checklists that provide current knowledge of names and their meanings. These checklists are typically at broad taxonomic levels and may be global or cover a particular geographical area. The process of evaluation and synthesis that goes into a taxonomic checklist can take many forms: some checklists are produced by single individuals (Coues 1873), while others are started by a single individual but are then completed or updated by later taxonomists (Ridgway and Friedmann 1901; Peters et al. 1931). Several contemporary checklists are produced, maintained and updated by governmental organizations, such as the Integrated Taxonomic Information System or ITIS (U.S. Geological Survey 2017b). Some checklists are published by an organization and produced and updated by a committee, such as the Check-List of North American

Birds. This checklist was first published by the American Ornithologists' Union (AOU) in 1886 along with a "Code of Nomenclature" that laid out how binomial and trinomial names should be used, how names should be defined and published, and how they may be synonymized (American Ornithologists' Union 1886). This checklist continues to be updated by the AOU's North American Classification Committee to the present day, with updates made annually (Chesser et al. 2017).

I focus on a few specific type of differences between checklists in my dissertation. I refer to any difference from one checklist to the next as a "change", whether it is the addition or deletion of a name or the modification of its circumscription. I refer to one particular type of change as a "correction": those in which the circumscription of an existing species is modified, whether by expanding it to include other species ("lumping") or dividing it into multiple species ("splitting"). Using the term "correction" is intended to connote that an existing entity is being amended while also serving as a reminder that each correction is being carried out in order to improve how accurately the checklist represents taxonomic knowledge at a particular point in time. The term "emendation" has a similar connotation; I chose "correction" as more aesthetically pleasing. Note that while the goal of corrections is to improve accuracy, there is no way to be certain that these changes will continue to be recognized by taxonomists in the long term – in fact, in Chapter 2. The tempo and mode of the taxonomic correction process in North American Birds over the last 127 years, I measure how often these

corrections themselves need to be corrected. Thus, corrections do not represent a completely accurate final state, but a correction on a path to eventual completeness.

Individual checklists provide summaries of taxonomic views at particular places and times, and need to be reconciled with each other to provide a single, continuous view of taxonomic change over time. By contrast, when a single checklist has been updated over time, such as ITIS or the AOU Checklist, the names can be assumed to have identical circumscriptions from one edition to the next unless shown otherwise, reducing the amount of taxonomic expertise necessary to stitch them together. Such stitching provides a historical record of taxonomic decisions, which in turn reflect changing patterns of taxonomic practice over time. The use of such a record can test hypotheses about tempo and mode of taxonomic changes. In Chapter 2. The tempo and mode of the taxonomic correction process in North American Birds over the last 127 years, I use the AOU checklists mentioned earlier to determine whether North American bird species, which appear to have been largely described, are still being corrected and whether there is any sign that ongoing corrections are slowing down. In particular, I am interested in whether North American birds follow the same pattern of increasing splitting that have been reported from other vertebrate groups (Isaac, Mallet, and Mace 2004) and that have been reported for birds globally (Sangster 2009, 2014), and if this increased splitting is the result of previously lumped species being resplit, essentially acting as a "re-correction" process as taxonomic corrections are themselves corrected over time.

Reconciling taxonomic checklists also has a practical aspect to it, as much biodiversity data has been identified using taxonomic checklists. All present checklists are likely wrong, but it is unclear in what ways they are wrong and so where they are likely to be corrected in the future. This has led some scientists to recommend that the taxonomic publications used to identify taxa be recorded with taxonomic identifications (Meier 2017). In Chapter 3. The components of the taxonomic discovery process and the effect of abundance in the ongoing discovery of North American freshwater algae, I attempt to determine what effect changes in taxonomic checklists may have on the interpretation of biodiversity data, and whether certain types of species (such as those in larger genera or with more observations) are more likely to be redescribed than others. These two checklists provide two very different perspectives on redescription: the North American bird checklist provides a historical perspective going back 127 years and focuses on a group in which taxonomic description has all but ceased, with only three new species described since 1950. In North American freshwater diatoms, the checklist only covers the last 15 years, new species description is ongoing, and the taxonomic changes directly affect an associated biodiversity dataset that uses the checklist as an index.

For the full value of taxonomic checklist series to be realized requires many more checklist series to be digitized, analyzed and shared. The source data for such analysis is easy to produce, if rarely produced in practice: maintainers of taxonomic checklists need only maintain archives of previous lists of recognized species. In

order to carry out the analyses in the first two chapters in my dissertation, I wrote a piece of software that can load a series of taxonomic checklists and identify additions and deletions among consecutive checklists. These can then be annotated into renames, lumps and splits, allowing synonymy and circumscription change rates to be identified. I describe this software in Chapter 4. SciNames: a tool for assembling datasets of taxonomic changes to improve biodiversity data reconciliation using taxon concepts, along with a few test cases that demonstrate its value in determining checklist change rates and aggregating biological data using these changes. While the software was developed with some specific use cases in mind based on my dissertation effort, it was designed with extensibility in mind in order to flexibly serve new uses as needed.

My dissertation focuses on taxonomic checklists as a centrally important product of taxonomy, and attempts to develop measures of the stability of checklists among critical dimensions: names, clusters of names that share the same meaning, and circumscriptions. I use these measures in order to understand how and why checklists change, and how these changes can affect the interpretation of biodiversity data. My goal is to test the use of taxonomic checklist series as a source of useful knowledge about both biodiversity and how our knowledge of biodiversity grows and changes with changes in philosophy and technology.

CHAPTER 2. THE TEMPO AND MODE OF THE TAXONOMIC CORRECTION PROCESS IN NORTH AMERICAN BIRDS OVER THE LAST 127 YEARS[1]

Abstract

While studies of taxonomy usually focus on species description, there is also a taxonomic correction process that retests and updates existing species circumscriptions on the basis of new evidence. These corrections may themselves be subsequently retested and recorrected. We studied the contribution of this correction process utilizing the *Check-List of North and Middle American Birds*, a well-known taxonomic checklist that spans 130 years. We identified 142 lumps and 95 splits across sixty-three versions of the *Check-List* and found that while lumping rates have markedly decreased since the 1970s, splitting rates are accelerating. We found that 74% of North American bird species recognized today have never been corrected (i.e., lumped or split) over the period of the checklist, while 16% have been corrected exactly once and 10% have been corrected twice or more. Since North American bird species are known to have been extensively lumped in the first half of the 20th century, we determined if most splits seen today are the result of those lumps being recorrected. 5% of lumps and 23% of splits fully reverted previous corrections, while a further 3% of lumps and 13% of splits are partial reversions.

[1] Currently in review in PLOS ONE with Denis Lepage (Bird Studies Canada, Port Rowan, Ontario, Canada) and Robert Guralnick (University of Florida, Gainesville, Florida) as co-authors, submitted July 6, 2017. The analyses were planned and executed by me.

These results show a taxonomic correction process with moderate levels of recorrection, particularly of previous lumps. However, 81% of corrections do not revert any previous corrections, suggesting that the majority result in novel circumscriptions not previously recognized by the *Check-List*. We could find no order or family with a significantly higher rate of correction than any other, but twenty-two genera do have elevated rates. Given the currently accelerating rate of splitting, prediction of the end-point of the taxonomic recorrection process is difficult, and many entirely new taxonomic concepts are still being, and likely will continue to be, proposed and further tested.

Introduction

The goal of taxonomy is to provide a complete, accurate catalogue of planetary biodiversity. When taxonomists encounter a putative new species, they collect evidence to support the hypothesis that it is distinct enough from any known species to necessitate its own name. If so, this species is formally described and is given a new name under the appropriate codes of nomenclature (Ride et al. 1999; McNeill et al. 2012). Over 16,000 species have been described every year between 2000 and 2010 (IISE 2011), and both the number of new descriptions and the number of authors involved in species description across multiple plant and animal groups have been rising since the 1750s, while the number of species described by each author has been falling (Tancoigne and Dubois 2013; Joppa, Roberts, and Pimm 2011). These observations may suggest that more taxonomists are chasing

fewer remaining species, and thus species description may be approaching completion (Costello, Wilson, and Houlding 2013). But the taxonomic process remains incomplete even after all species have been initially described: an unknown proportion of described species will eventually be re-tested and, if falsified, may be rejected in favor of other hypotheses of conspecificity (Sluys 2013). The proportion of species hypotheses that will eventually be falsified may be expected to vary over time as techniques and species delimitation philosophies change and as more evidence accumulates. Understanding how often these corrections take place may allow us to estimate when all taxonomic work — both species description and taxonomic corrections — will finally be completed.

Taxonomic corrections also have a practical impact on lists of recognized species, widely used in biological analyses and often treated as stable despite ongoing corrections (Padial and de la Riva 2006). In particular, there has been a sharp increase in the number of subspecies being raised to full species across a wide range of animal groups in the last few decades (Agapow et al. 2004), including primates (Isaac, Mallet, and Mace 2004; Groves 2014), amphibians (Padial and de la Riva 2006), bovids (Heller et al. 2013) and birds (Sangster 2009). This phenomenon, termed "taxonomic inflation" by Isaac *et al.* (Isaac, Mallet, and Mace 2004), does not yet have a widely-accepted explanation. Some scientists treat it as the result of a shift in taxonomic practice, either from the biological species concept to the phylogenetic species concept (Isaac, Mallet, and Mace 2004) or from an assumption of free interbreeding to an assumption of reproductive isolation (Gill 2014), and

have suggested that the number of globally recognized bird species may double as a result of this change in criteria (Barrowclough et al. 2016). Other scientists point out that the increase in the number of subspecies being raised to species began in amphibians in the 1950s (Padial and de la Riva 2006), several decades before the phylogenetic species concept was proposed (Cracraft 1983). They further point to studies of the global bird taxonomic literature that have shown that, in practice, diagnosability rather than reproductive isolation has remained the most commonly used criterion to justify proposed taxonomic changes between 1950 and 2009, regardless of the underlying species concept being used (Sangster 2009, 2014). Understanding how widespread these corrections are, how often a scientific name is likely to be corrected, and whether this number varies by taxonomic groups such as orders, families and genera may allow us to determine if there are particularly stable taxonomic groups and may provide means to predict where corrections are likely to be made in the future.

Taxonomic corrections may themselves require correction, which will then fuel further taxonomic inflation. Remsen Jr noted in 2015 (Remsen Jr. 2015) that "virtually all current systematists, regardless of species concepts, recognize that current species limits in many bird groups are far too broad, incorrect, or weakly justified", and posited that "overapplication of Biological Species Concept (BSC) criteria by many taxonomists in the mid-20th century, often without explicit rationale, demoted by mere pen strokes hundreds of taxa from the rank of species to subspecies, before the importance of vocal differences was recognized". Some

systematists in the 1920s and 1930s were equally skeptical about demoting species to subspecies (Ridgway 1923; Swarth 1931; Stone 1935; Grinnell 1935). This all points to a current taxonomic recorrection process, in which corrections made in the first half of the 20th century are now being reverted in light of new evidence and better tools. Quantifying the contribution of this process to building the list of currently recognized species may provide a means to extrapolate what proportion of species circumscriptions generated before 1980 are likely to eventually be reverted. More broadly, it shines a light on the trajectory of all forms of taxonomic correction: by understanding which species are corrected and recorrected, we gain a deeper understanding into how the correction process progresses and how long it takes to complete. We delineate more focused, testable questions below, but first discuss the importance of checklists for examining taxonomic corrections over long periods of time.

The importance of taxonomic checklists

Taxonomic corrections are published in a wide variety of scientific literature, from scientific monographs to taxonomic checklists to general-interest identification guides. Previous analyses have surveyed a set of journals where taxonomic corrections are likely to be published (e.g. Sangster 2009, 2014), but there is no easy way to determine if a particular proposal has gained widespread recognition within its taxonomic community or is considered a purely speculative opinion. Conventional methods to gauge the impact of a publication, such as citations counts, do not help: a contentious proposal may be heavily cited by scientists

disputing it, while a generally accepted proposal may only be cited a few times before being incorporated into compiled resources, which may then be cited instead.

One source of taxonomic corrections representative within a taxonomic group and generally recognized by both taxonomists and other biologists are taxonomic checklists. These are expert-curated authoritative lists of recognized species within a taxonomic group in a particular geographical area. Checklists are neither universally used nor necessarily congruent: different biologists often disagree on which taxonomic checklists they use when identifying taxa, and checklists may circumscribe species differently on the basis of differences in available evidence, taxonomic philosophy or tools used (Lepage, Vaidya, and Guralnick 2014). Taxonomic checklists may be critiqued by taxonomists (Remsen Jr. 2015; Heller et al. 2013) and have been used to estimate the stability of binomial names (Olson 1987; Rising and Schueler 1972). In this study, we focused on one such checklist, which has been maintained over the last 130 years by the North American Classification Committee of the American Ornithologists' Union (AOU): the *Check-List of North American Birds*, hereafter referred to as the "AOU Checklist". This checklist was first published in 1886, and since then has been updated in six major and fifty-seven minor updates through 2016 (Chesser et al. 2016). The North American Classification Committee reviews corrections submitted to it based on changes proposed in the literature, and accepts those supported by two-thirds of its members (The American Ornithologists' Union 2017). These corrections are then published as a series of editions and supplements. The first update was published in

1889, giving us 127 years of corrections until 2016. The last complete edition (the 7th edition) was published in 1998 (American Ornithologists' Union 1998). Supplements have been published at an average of one every 2.03 years. Since 2002, updates have been published every year (see S1 Table).

The AOU Checklist therefore provides a community review process for taxonomic corrections. It continues to be widely used as an authoritative source for taxonomic names among both professional ornithologists and an often highly engaged public, the birding community, either directly or indirectly through birding organizations and field guides that track the AOU Checklist. These include the National Audubon Society's Bird Guide App (National Audubon Society 2017), the Cornell Lab of Ornithology's eBird/Clements Checklist (Schulenberg and Iliff 2014), the American Birding Association Checklist (Swick 2016), and the Sibley Guide to Birds (Sibley 2012). Species description in North American birds is largely considered to be close to completion (Bebber et al. 2007) after over 250 years of study (Catesby 1731), but the number of North and Middle American bird species is increasing rapidly as previously described species are being added to it. The AOU Checklist has grown from approx. 1,908 species in 1983 (American Ornithologists' Union 1983) to 2,127 species in 2016 (Chesser et al. 2016), an 11.5% increase within a consistent geographical area. Since birds have been central to the development of the biological species concept (Mayr 1942), the phylogenetic species concept (Cracraft 1983), as well as Remsen Jr observations of past, potentially problematic

corrections, they are a particularly apt group to begin studies of taxonomic correction and recorrection processes.

Key questions

Our work here focusses on corrections that alter the circumscription of a scientific name without altering the name itself. These are of two kinds: the division of putative species into multiple species ("splits"), which usually occurs through the raising of a subspecies to a full species, and the union of putative species into a single species ("lumps"). In order to understand how taxonomic circumscriptions change post-description, we quantify several rates. We define the "correction rate" as the proportion of currently recognized species that have ever been corrected, and the "recorrection rate" as the proportion of currently recognized species that have been corrected more than once. The "full reversion rate" is the proportion of all corrections that completely reverted an earlier correction (i.e. when a lump is subsequently resplit, or a split is subsequently relumped). Note that full reversions may not yield exactly the same circumscriptions. We further define a more general "reversion rate" as the proportion of all corrections that have been partially or completely reverted, in which two or more split species are relumped or where two or more lumped species are resplit, along with other sister species. To quantify how these taxonomic correction lead to the current taxonomy, we summarized the sequence of lumps and splits that led to each of the currently recognized species.

To test whether newly recognized bird species were the result of resplitting previous lumps, we first determined the proportion of all splits that were the result

of a previous lump and then tested whether lumps were as likely to be reverted as splits were. If this period of splitting is largely the result of undoing lumping from before 1980, we would expect to see many more splits reverting previous lumps than vice versa. If, on the other hand, most splits are unconnected with previous lumps, this suggests taxonomists are generating novel circumscriptions and not solely correcting a backlog of incorrect lumping. We also ask if certain bird groups, at multiple taxonomic hierarchical levels, are more likely to be corrected than others, given that traits that make species delimitation more difficult may be shared among closely related species. For instance, some traits may make species boundaries more difficult to identify or by making the species themselves harder to study. Our analyses thus provide insight into past and current taxonomic correction processes for North American birds, especially how often entirely new concepts have been and are still forming as opposed to the re-recognition of previously subsumed concepts.

Materials and Methods

Source data

The AOU Checklist consists of sixty-four checklists published between 1886 and 2016: seven major editions, which list every recognized species, and fifty-seven "supplements", which list changes to the checklist since the previous supplement (S1 Table). We began with lists of additions, deletions and changes in scientific names to the AOU Checklist collected by one of the authors (DL) for checklists published between 1886 and 2012. These changes were collected as part of the

online database Avibase (Lepage 2017), which also contains information on which circumscriptions are entirely contained within others (Lepage, Vaidya, and Guralnick 2014). Based on this information, we excluded additions and deletions that did not involve overlapping circumscriptions – in most cases, these were the results of changes in distributional records, such as when a previously described species was discovered in North America. We checked changes involving overlapping circumscriptions against the AOU Checklists themselves to identify those that were explicitly stated to be a lump or split in the publications; for instance, "...we divide *B[ranta] canadensis* by recognizing a set of smaller-bodied forms as the species *B. hutchinsii*..." from the 45th supplement (Banks et al. 2004). Lumps or splits identified by Avibase were excluded from our analyses if the AOU Checklist did not explicitly indicate them as such, since Avibase may have made this determination based on the view of later taxonomists while we aimed to capture the contemporary view as far as possible. As a result, our measures are conservative counts that are likely smaller than the true values – a more thorough study of the contemporary literature might lead to evidence that a particular addition was known at the time to be a split. Since the 34th Supplement provided a list of all species recognized in 1982 and the AOU published an online spreadsheet of recognized species in 2016, we used these to correct any discrepancies that may have entered our dataset before those dates. For checklists between 2013 and 2016, which postdate our initial export of Avibase data, we extracted the lumps, splits and name changes directly from the supplements themselves (Chesser et al. 2013, 2014,

2015, 2016). In all, we found 148 lumps and 191 splits recognized by the AOU Checklist between 1889 and 2016, covering North America excluding Hawaii before 1982, and covering North and Central America including Hawaii after 1982.

Our analysis was complicated by a large increase in the geographic range of the AOU Checklist in 1982 and 1983, expanding to include Mexico, the Hawaiian Islands, the Caribbean Islands and Central America while removing species found only in Greenland. From approx. 858 species recognized in the 33rd Supplement (Eisenmann et al. 1976), the number of recognized species rose to 937 species in the 34th Supplement (Eisenmann et al. 1982) and to approx. 1,908 species in the 6th Edition (American Ornithologists' Union 1983) (S1 Table). To obtain a consistent picture of taxonomic corrections over as long a time period as possible, we eliminated all additions, deletions, renames, lumps and splits involving species first added to the checklist after 1981, thus isolating corrections among species in continental North America. This resulted in 142 unambiguous lumps and 95 unambiguous splits recognized by the AOU Checklist between 1889 and 2016 (S2 Table). After eliminating these changes, the number of recognized species varied from 771 (in 1886) to 875 (in 1956), before reaching a final count of 851 species in 2016 (S3 Table). Of these 851 species, 17 were the result of "extralimital" lumps and splits that took place outside of the AOU Checklist's geographical area, resulting in 834 currently recognized species after filtering. We eliminated ten checklists because no unambiguous lumps or splits took place in them (1894, 1909, 1912, 1920, 1957, 1983, 1991, 1998 and 2009). We calculated the cumulative change in the

number of lumps and splits over the last 127 years (Figure 1) and summarized these changes by decade to look at overall trends (Figure 2).

To account for synonymy while measuring these rates, we assembled "name clusters" that link together species names that have been renamed. For example, *Phyllopseustes borealis* was first added to the AOU Checklist in 1886, but has since become known as *Acanthopneuste borealis* and *Phylloscopus borealis* as it was moved between different genera. These three names constitute a single name cluster, and a lump involving one name will be matched in our analysis with a split involving another name in the same name cluster. All 834 name clusters are included in S3 Table, where extralimital name clusters are indicated by an 'NA' in the 'Order' column.

Taxonomic corrections

To measure how often individual lumps and splits are reverted, we identified partial and full reversions for every lump and split. A full reversion is one where the other change exactly undoes the first one, such as *Gallinula galeata* being lumped into *Gallinula chloropus* in the 18th Supplement (Stone et al. 1923) but then resplit in the 52nd Supplement (Chesser et al. 2011). A partial reversion is where two or more lumped species are resplit or two or more split species are relumped along with other sister species. An example is *Rallus obsoletus* being lumped into *Rallus longirostris* in the 19th Supplement (Wetmore et al. 1944), but later resplit in the 55th Supplement (Chesser et al. 2014) into *R. obsoletus* and *R. crepitans*. It is possible but not guaranteed that the circumscription for *R. obsoletus*

as of the 55th Supplement is congruent to the circumscription for *R. obsoletus* before the 19th Supplement; therefore, our analysis assumes that every lump or split results in a new circumscription. The full list of reversions is included in the table of lumps and splits (S2 Table). To test whether resplitting previously lumped species directly caused increases in recognized species, we determined whether lumps were as likely to be resplit as splits were to be relumped.

For each currently recognized species name cluster, we identified the sequence of lumps and splits in which they have been involved. In particular, we wanted to know what proportion of name clusters had never been corrected, what proportion had been corrected one or more times (the "correction rate"), and what proportion had been corrected more than once (the "recorrection rate"). In order to determine the trajectory of corrections necessary to obtain the current name cluster, we tallied up the number of lumps and splits each name cluster had been involved with in chronological order. We also counted the total number of lumps and splits for each name cluster. Since every lump and split potentially results in a new circumscription (i.e. a new taxon concept *sensu* Franz *et al.* (N. Franz, Peet, and Weakley 2008)), this gives us the number of circumscriptions associated with each species name cluster. This is included in the table of name clusters (S3 Table).

Differences in correction rates among higher-level taxa

To determine whether different taxonomic groups showed significantly different correction rates, we modeled the number of taxonomic corrections (lumps + splits) involving currently recognized name clusters as a Poisson distribution, in

which the rate at which new corrections are made to species (λ) is assumed to be constant within a taxonomic group. Since our analysis focuses on 834 currently recognized species clusters, we used the higher taxonomic system provided by the AOU Checklist in 2016. Our model had three hierarchical levels of grouping: at the level of order (π), family (τ) and genus (ρ). Additionally, we included an offset to account for the different lengths of time that different species have been in the checklist. Our hierarchical model can be described as:

$$y_i \sim Poisson(\lambda_i)$$

$$log(\lambda_i) = \lambda_0 + \pi_i + \tau_{j[i]} + \rho_{k[j[i]]} + log(t_i)$$

Each of these parameters were modeled as normally distributed random variables, with a mean of zero and with variable standard deviations (σ_π, σ_τ and σ_ρ respectively). t_i is the number of checklists that this species has been recognized in the AOU Checklist, to control for some species having been recognized by the AOU Checklist earlier, giving them a longer time span within which to be lumped or split than others. This model failed to converge in rSTAN 2.15.1 (Stan Development Team 2017), and so we used transformed parameters to define standard normal deviations that were multiplied by the variable standard deviations (see S7 Code). This model converged successfully in rSTAN and gave us an estimate of the overall mean rate of correction (λ) as well as the mean rate for every genus, family and order (S4-S6 Tables).

Results

Overall trends in lumping and splitting

As of 2016, the AOU Checklist recognizes 2,127 species from North and Central America, including Hawaii (Chesser et al. 2016). The rate of species description among these species has been falling steadily: 191 species (9%) have been described since the AOU Checklist was first published in 1886, half of which (101 species or 4.8%) have been described since 1900, and only 14 species (0.7%) have been described since 1950. When we looked at the 834 species remaining in our checklist after filtering out names added after 1981 as well as extralimital species, 30 (3.6%) were described since 1886, 15 (1.8%) since 1900 and only three species (0.4%) since 1950. Thus, primary species description in this group appears to be proceeding at a very low but non-zero rate.

In contrast, taxonomic corrections have been proceeding at a rapid rate: we discovered 142 unambiguous lumps and 95 unambiguous splits on species name clusters added before 1982. Examining the cadence of lumping and splitting (Fig 1), we note large numbers of lumps, in particular the 40 lumps in the 4th edition (American Ornithologists' Union 1931), 30 lumps in the 19th supplement (Wetmore et al. 1944), and 16 lumps in the 32nd supplement (Eisenmann et al. 1973). While there are no specific spikes in the number of splits, most of the splits (70, or 73.7%) in our dataset took place in or after 1980. Cumulative plots show that lumping has all but ceased since 1980, while splitting rates have sharply increased since the 1980s and continue to accelerate to the present day (Figure 2). Based on the trends

in the data, new formation of taxonomic concepts in North American birds since 1950 and particularly since 1980 is mainly driven by splitting of taxa. As noted by several authors (Gill 2014; Barrowclough et al. 2016), the era of splitting appears to be far from over.

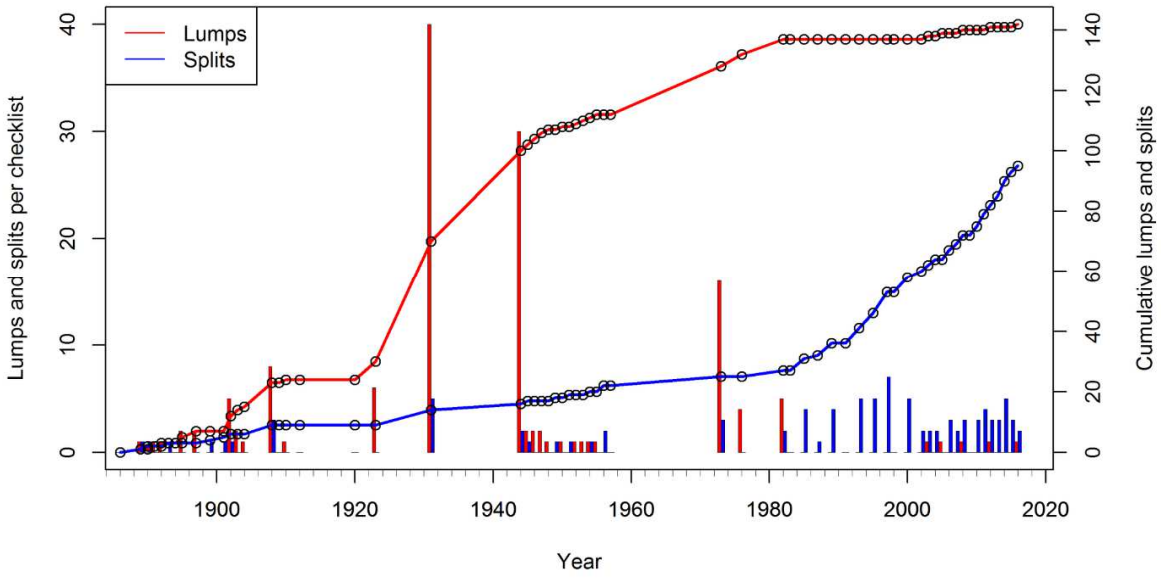


Figure 1. Individual and cumulative lumps and splits within the AOU Checklist between 1886 and 2016. *Each circle represents a single checklist, showing periods of activity (1944-1957, 1980-2016) as well as periods of relative inactivity (1920s and 1960s).*

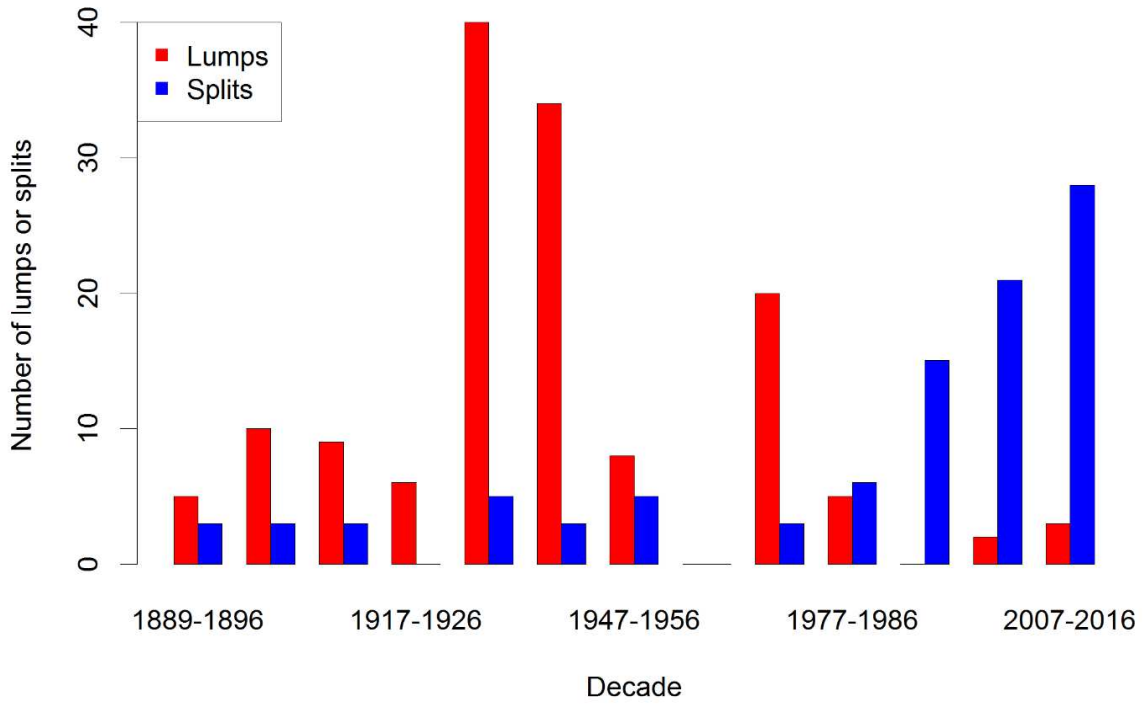


Figure 2. Bar plots of number of lumps and splits by decade showing accelerating number of splits per decade in the present. *Note that the first decade is incomplete, as we only have data on the eight years from 1889 to 1896.*

Full and partial reversions

We begin by considering the corrections themselves to determine the scope of original correction and subsequent recorrection. We found a total of 142 lumps and 95 splits occurring amongst currently recognized species that were first added to the AOU Checklist before 1982. Of these, 7 lumps (4.9%) and 22 splits (23.2%) fully revert a previous split or lump, respectively, for an overall reversion rate of 12.2%. If we count both full and partial reversions, these numbers increase to 12 lumps (8.5%) and 34 splits (35.8%) partially reverting an earlier correction, for an overall partial reversion rate of 19.4%. Thus, 80.6% of all corrections do not revert a

previous correction, and 64.2% of splits do not revert a previous lump. There were significantly more splits than lumps both fully reverting previous corrections (exact binomial test, $p < 0.01$) as well as partial corrections (exact binomial test, $p < 0.01$). We found the proportion of splits reverting previous lumps were significantly higher than would be expected based on the ratio of lumps to splits in our dataset (Fisher's exact test, $p < 0.001$). Less than half of all lumps have been partially (36 lumps, 25.4%) or fully (22 lumps, 15.5%) reverted, suggesting that the resplitting process is either mostly incomplete or that most lumps may never be resplit.

We can also determine the proportion of all corrections involved in any recorrection, either by correcting a previous correction or by being corrected in the future. We found 54 corrections (22.8%) involved in full reversions while 86 corrections (36.3%) were involved in partial reversions. Therefore, 63.7% of all corrections are neither correcting a previous correction nor have yet been corrected by a future correction.

Corrections involving currently recognized species

Identifying the species affected by the corrections we have catalogued is complex: every correction affects multiple species, and species that are lumped are no longer recognized as species by the AOU Checklist. Species may also be removed from the AOU Checklist if the species is no longer found within the checklist area, or added not for any taxonomic reason but solely because it has been introduced into the checklist area. Thus, there is no clear denominator of the total number of

species recognized with which we can compare the number of species affected by taxonomic corrections.

Instead, we focused our analysis on one particular question: if a researcher today were to use a species name currently recognized by the AOU Checklist, how likely is this to be a species that has been corrected within the lifetime of the Checklist? As previously described, to maximize the time period we could cover, we started with the 2,127 species currently recognized, eliminated species added after 1981 and obtained 834 currently recognized species names (Table S3). Of these, 615 species (73.7%) have never been corrected in the course of the Checklist (Fig 2), suggesting that most species are not corrected over long periods of time.

To determine the sequence of lumps and splits affecting each species, we identified all lumps and splits involving the species (as either source or result) and arranged them in chronological order. Fewer than 2.2% of species were involved in more than two corrections, and so we have summarized these results on the basis of the first two corrections involving each species. Of the 219 species (26.3%) that have been corrected one or more times, more species were first lumped (129 or 58.9%) than first split (90 or 41.1%). As a reminder, these are the number of *species* that are involved in lumps and splits, not the number of corrections themselves. However, 43.4% of species involved in a lump were subsequently involved in a split, while only 16.7% of species involved in a split were subsequently involved in a lump. 85 species (10.2%) were corrected two or more times. Thus, the overall correction rate was 26.3% and the overall recorrection rate was 10.2%.

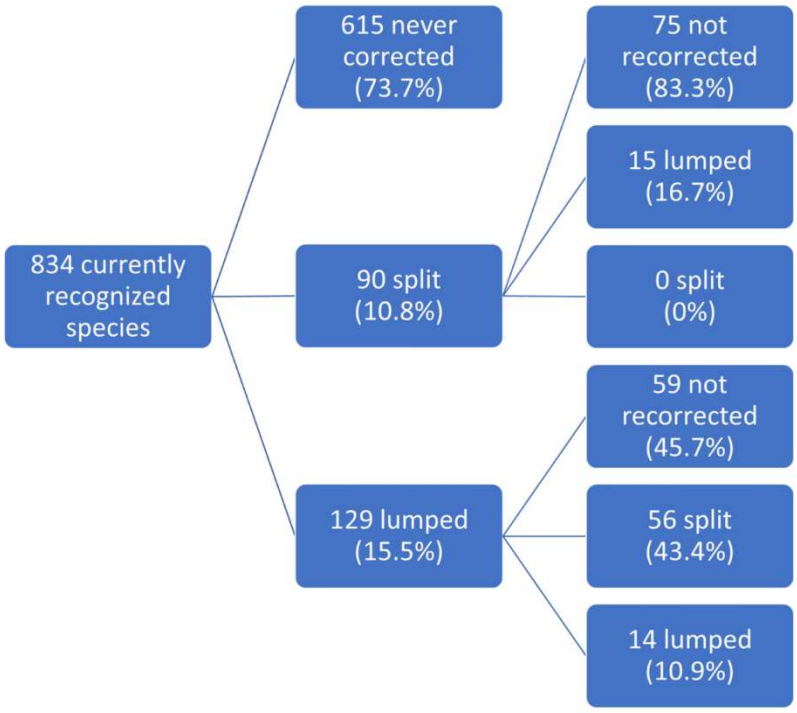


Figure 3. A diagrammatic representation of the corrections involved in generating the 834 currently recognized name clusters. *Note that a lump followed by a split does not imply that the split reverted the lump; different species might have been split out of the lumped circumscription to obtain the current circumscription. We see relatively low rates of initial corrections, but once corrected, 43% of species involved in lumps are later involved in splits, while only 17% of species involved in splits are subsequently involved in lumps. 18 species that were involved in more than two corrections are summarized by their first two corrections above.*

Which species are most likely to be lumped or split?

We used a Bayesian hierarchical model to determine if some orders, families or genera were more or less likely to be associated with multiple taxon concepts than others among the 834 species we used in our analysis. We used the contemporary taxonomy used by the AOU Checklist in 2016. Our model fit a Poisson distribution with $\lambda = 0.3985$. While no orders or families showed significantly higher or lower rates of correction, 22 genera showed significantly higher rates of corrections: *Ammodramus*, *Anser*, *Aphelocoma*, *Artemisiospiza*,

Baeolophus, *Branta*, *Butorides*, *Dendragapus*, *Empidonax*, *Gallinago*, *Gallinula*, *Junco*, *Leucosticte*, *Limnodromus*, *Melanitta*, *Melozone, Puffinus*, *Quiscalus*, *Rallus*, *Sternula*, *Sula*, and *Troglodytes* (S4-S6 Table). These correspond to 6.5% of the 338 genera in our dataset, and belong to fifteen families across eight orders.

Discussion

Birds are often cited as a taxon in which species description is likely to be complete – for example, Bebber et al. (2007) estimated on the basis of species description curves that only 26-93 bird species remained to be described. The AOU Checklist supports this pattern, with over 90% of currently recognized species having been described before the Checklist was first published in 1886, and a mere fourteen species described since 1950. This does not represent a stable taxonomic end-state, however. When only considering species added before 1982 to the American Ornithological Union checklist, i.e. those species that was recognized by the checklist when it was limited to North America excluding Mexico, we found 142 lumps and 95 splits which were involved in the correction of 218 currently recognized North American species (correction rate: 26.3%), of which 85 currently recognized species names (recorrection rate: 10.2%) were involved in more than once correction.

We did not find a concentration of corrections in any one order or family, but 6.5% of North American bird genera in our study showed significantly higher rates of taxonomic correction. The lack of a higher taxonomic signal, related to shared

characteristics and life-history, and no immediately obvious other factor such as size of the genera, suggest that these higher rates may be due to historical reasons. We note however that these numbers only reflect a part of the complete debate over these circumscriptions, since we analyze changes within a single checklist. Thus, a species circumscription that is heavily debated in the literature may not have been recognized by the AOU Checklist until they felt that the evidence overwhelmingly supported one interpretation. An example of this is the species *Branta hutchinsii*, which had been recognized as a subspecies of *Branta canadensis* by the AOU Checklist until it was raised to a full species in the 45th Supplement (Banks et al. 2004). Before the AOU Checklist was first published, both its original author (Swainson and Richardson 1831) and John James Audubon (Audubon 1835) treated it as a separate species, and proposals for treating it as a separate species date back until at least 1946 (Aldrich 1946). Thus, we re-emphasize that both the per-genus correction rates and the overall correction, recorrection and reversion rates we document reflect a conservative measure of all proposed corrections in the literature, but are likely accurate for the widely-recognized corrections that scientists use in practice.

Our results show a clear period of lumping in the 1920s to the 1980s, followed by a period of rapid splitting, at least in the AOU checklist. The lumping period coincides with the ascendance of the biological species concept (BSC) in the late 1930s, and the surge in splitting generally coincides with the formalization of the phylogenetic species concept (PSC) in the 1980s. 19.4% of all lumps and splits in our

dataset are full or partial reversions of a previous correction, 74% of which are splits reverting a previous lump. Reversions are clearly a part of the current period of splitting, but the majority (64.2%) of splits do not partially or fully revert a previous lump. Furthermore, 80.6% of all corrections do not partially or fully revert a previous correction, showing that the generation of circumscriptions novel to the AOU Checklist have been and continue to be a critical part of taxonomic revision. Both previously uncorrected species as well as previously recognized corrections are being actively retested and corrected by North American bird taxonomists today.

Is the current era of splitting the result of a change in the philosophy (such as in the species concepts or the species delimitation criteria being used) or the availability of evidence (through better sampling and improved techniques)? This is a difficult question to answer, especially since conceptual advances in species delimitation are not decoupled from the availability of new tools and methods. We document a clear shift from lumping to splitting in the 1980s, followed by an accelerating rate of splitting into the present day, and this timing may seem to support the interpretation that this shift is the result of a change in the species concept being used by taxonomists. However, there is little evidence that such a shift explicitly took place within this community: for example, the AOU Checklist's authors "strongly and unanimously continues to endorse the biological species concept (BSC)" in 1998 (American Ornithologists' Union. Committee on Classification and Nomenclature. 1998). However, this does not mean that taxonomists might not have begun to use tenets of the phylogenetic species concept

(PSC) when delineating species. Sangster's bibliometric analysis (Sangster 2014), while not focusing exclusively on checklists, provides useful evidence here: he found that the majority of lumps and splits proposed for global bird species between 1950 and 2009 used diagnosability as a criterion for delimiting species, with reproductive isolation used in fewer than half the proposals in every decade except the 1970s, when it briefly reached 50%. Coincident have been development of concepts such as the Comprehensive Biological Species Concept in 1999 (Johnson, Remsen Jr, and Cicero 1999), which advocates for a less narrow interpretation of the BSC. Thus, whether or not there has been an explicit shift to the PSC, there may have been an implicit change in taxonomic practice as a result of its development that has led to the patterns we see in our paper.

The 1980s were also a period of great technological innovation in both biology, with the development of Sanger sequencing (1977) and the polymerase chain reaction (1983), and in the world at large, with the development of the personal computer (1977, 1981) and NSFNET, the predecessor of the Internet (1985). Any of these, as well as any number of changes in the funding or production of taxonomic work, may have led to an increased output from taxonomists, which we would observe as an increased rate of correction since the 1980s. We observe that rates of species description (Tancoigne and Dubois 2013; Joppa, Roberts, and Pimm 2011) as well as the number of scientists involved in species description (Costello, Wilson, and Houlding 2012) have been increasing since the 1950s. Whatever factors are responsible for that increase may also be increasing the number of taxonomists

testing and correcting taxonomic circumscriptions, leading to the accelerating splitting rates we see. Further, some of that work appears to have been put into the recorrection of previously corrected species circumscriptions.

Extrapolating this pattern into the future and using taxon concepts (sensu Franz, Peet, and Weakley 2008) as the key unit, rather than simply the species names, we expect a continuing period in which both the development of entirely new concepts and the reversion of previously recognized concepts are carried out side-by-side. The refinement of theoretical approaches to species delimitation and growth in empirical datasets such as genomic data should lead to improving species circumscriptions and to fewer taxonomic errors remaining to be found and fewer taxonomic debates that remain to be conclusively settled. Based on this, we can extrapolate a taxonomic end state in which taxonomic corrections fall to a low, but non-zero rate, in much the same way species description rates have in North American birds. This rate will never reach exactly zero, not only because new evidence will continue to refine our view of historical speciation, but also because speciation is an ongoing process that will continue to lead to divergent lineages and thus to new species, likely at a very low rate. Species description and lumping appear today to be proceeding at these low but non-zero rates, especially considering the much higher rates they demonstrated in the 1800s and between 1930 to 1960 respectively. By comparison, splitting is proceeding at an unprecedented rate within the checklist, which continues to accelerate. If they predominantly reverted previous lumps, we might have been able to extrapolate

when all previous lumps might be fully resplit, but we find that only 25% of lumps have been reverted, and 81% of all changes do not revert a previous change. Therefore, our results do not provide an empirical means to predict when this end state might be reached. However, we do note that continuing acceleration along the trajectory we show here could hasten what others (Gill 2014) have argued is likely to be a slow process.

How general are the patterns we show here for other taxa and regions? Bird taxonomy was strongly impacted by extensive lumping from the 1920s to the 1980s, but we still find that the outcome of splitting is as much new taxonomic circumscriptions as it is reversions to previously recognized circumscriptions. Among other groups in which taxonomic inflation has been observed, such as amphibians, primates and bovids, we might expect to see a similar pattern of mixed taxonomic corrections and recorrections explaining the increase in the number of recognized species. More broadly and across a larger spectrum of the tree of life, we still know little about groups where current description rates far swamp any taxonomic corrections. As studies like ours are replicated, we hope that broader answers to questions about the tempo, mode and potential end-states of taxonomic discoveries can be found.

A final motivation for our work was the extent to which taxonomic correction leads to errors when biodiversity analyses use species name without considering the different circumscriptions that may be associated with that name. In our dataset, we find that 74% of species were only associated with a single circumscription, 16%

of species were associated with exactly two circumscriptions (by being corrected once) and only 10% of species were associated with more than two circumscriptions (by being corrected two or more times). Thus, a still significant proportion of species have multiple taxon concepts that make simple taxon labels ambiguous. Errors may be minimized by focusing analysis on species known to have no taxonomic corrections, but in North American birds, no single order or family was found to be particularly unstable. This suggests one simply cannot avoid "problem-areas" in North American bird groups except possibly at the generic level. Instead, any broad-scale analysis that ignores taxon concepts is likely to introduce some error.

Our work draws attention to the parts of the taxonomic process that are often overlooked when focusing exclusively on species description and on names without reference to circumscriptions. Large public databases of species descriptions have been published by several organizations, including the Catalogue of Life ("Catalogue of Life" 2017), Zoological Record ("Zoological Record" 2017), the Plazi Treatment Bank (Miller et al. 2015) and downstream databases such as BioNames (Page 2013). These resources have facilitated many studies of the cadence of description patterns (Tancoigne and Dubois 2013), changing properties of species descriptions (Sangster and Luksenburg 2015) and estimations of the number of species remaining to be discovered (Costello, Wilson, and Houlding 2012). The first databases of circumscriptions have been built, including Avibase, which formed the basis of this study (Lepage 2017; Weakley 2015), and some biodiversity databases now incorporate circumscriptions, including citizen science platforms such as

iNaturalist (California Academy of Sciences 2017). New philosophical, ontological and software tools to identify (Cui et al. 2016), describe (N.M. Franz and Peet 2009), share (Taxonomic Names and Concepts Interest Group 2006; Laurenne et al. 2014) and reason over (M. Chen et al. 2014) taxonomic circumscriptions have become available recently, which we hope will lead to better, shareable circumscription datasets that provide a means to move beyond simply capturing name strings and towards the more fundamental units of biodiversity. The circumscriptions we used in this project and the corrections we based them on are only one interpretation of the taxonomic acts that we have studied; by making the data we used in this project available, we hope that future work will be able to build on our work to assemble larger datasets, leading to a much more thorough understanding of how taxonomic corrections have refined our knowledge of global biodiversity and how they will continue to do so in the future.

Supplementary Materials

Data tables containing information on the lumps and splits identified and used in these analyses as well as a list of currently recognized species after filtering out post-1982 additions are available on Figshare at https://figshare.com/s/99683d5f17fa4488a585. Upon successful publication of this chapter, those data will be made publicly available and given a DOI.

The data included in the supplementary materials include:

1. **S1 Table. List of AOU Checklist updates with authors and estimated counts of recognized species.**
2. **S2 Table. List of 142 lumps and 95 splits after filtering out all changes after 1981.** Includes information on all the changes that revert a particular change, as well as the subset of those reversions that are complete – where one change perfectly undoes another change. Note that "reversion" does not imply a particular ordering in time: both the initial change and all its partial or complete reversions will list the other change as reversions.
3. **S3 Table. 851 currently recognized species after filtering out all changes after 1981, including 17 extralimital species.** Includes a count and list of taxon concepts associated with each name, the 'trajectory' of changes (the sequence of additions, deletions, renames, lumps and splits) we know about associated with this name or its synonyms and which dataset this name and its synonyms were first added in. The remaining columns are from the 2016 Checklist of North and Middle American Birds, downloaded from http://checklist.aou.org on October 3, 2016. Extralimital species, i.e. those involved in lumps and splits but not found within the geographical area of the checklist, have 'NA' in all higher taxonomy columns and were not present in the 2016 Checklist.
4. **S4-S6 Table. Results of the hierarchical model at (respectively) the order, family and genus levels.** The total and mean number of redescriptions observed in that group are indicated. The 'min', 'max' and 'interval_width' values refer to the 95% credible interval around the 'mean' for the log difference in the λ

attributable to that group. The lower interval is greater than zero where the taxon has a significantly higher rate of taxonomic redescription than other groups.

5. **S7 Code. Raw data and analysis scripts for this project.** This raw data and analysis code is also available online at http://github.com/gaurav/aou_checklists and will be published on Figshare upon publication.

CHAPTER 3. THE COMPONENTS OF THE TAXONOMIC DISCOVERY PROCESS AND THE EFFECT OF ABUNDANCE IN THE ONGOING DISCOVERY OF NORTH AMERICAN FRESHWATER ALGAE

Introduction

Taxonomic names serve as an index to all biological knowledge: journal articles, textbooks, biodiversity databases and codes of law all use them as shared identifiers for biological entities. Description and recognition of these names are controlled by nomenclatural codes (Jach 2000; McNeill et al. 2012; Lapage et al. 1992), but these codes explicitly control only which names are considered valid, not how each taxon is defined. These definitions ("taxonomic circumscriptions") are proposed when the taxon is first described, but are hypotheses that may be subsequently updated as a result of new discoveries (Sluys 2013). Quantifying the overall process of taxonomic discovery is complicated by the number of interdependent processes that create taxonomic names, invest them with well-defined circumscriptions, and then update those circumscriptions over time. These updates improve our knowledge of planetary biodiversity, but do so at a cost to end-users who rely on stable taxonomic names: when names are unstable, the relationships between their circumscriptions need to be documented in order to ensure that data from multiple sources are aggregated correctly (Berendsohn 1995; M. Chen et al. 2014; J. Kennedy et al. 2006; N. Franz, Peet, and Weakley 2008). Here, I set out to quantify how quickly taxonomic discovery is occurring as recorded

by changing taxonomic checklists, investigate the effect of abundance on taxonomic redescription, and determine whether redescription could significantly affect the interpretation of biodiversity data.

The process of taxonomic description has been well-studied. Over 16,000 new animal and plant species were described every year between 2000 and 2010 (IISE 2011), and both species description rates and the number of authors involved in species descriptions have been rising steadily since the 1950s (Tancoigne and Dubois 2013; Joppa, Roberts, and Pimm 2011). Studies of taxonomic description have been used to identify sampling biases, such as the discovery that carnivore and primate species with larger geographical ranges tend to be described sooner (Collen, Purvis, and Gittleman 2004). Original descriptions can also be used to identify undersampled areas where new species may be found (ter Steege et al. 2016; Jones et al. 2009). The rate at which new species are discovered can be measured as a function of time, which has been used to estimate the number of species remaining to be described (Mora et al. 2011; Bebber et al. 2007), or as a function of the number of individuals sampled, which can be used to estimate how many new species may remain to be discovered through increased sampling in particular areas (Shen, Chao, and Lin 2003).

After taxa have been originally described, they may be redescribed on the basis of new evidence. Redescription is often simplified to a process of nomenclatural or taxonomic synonymization, in which one name (such as *Nitzschia umbonata*, a diatom species) is determined to be the junior synonym of another

name (such as *Nitzschia stagnorum*, see Figure 4). Use of the former name can generally be replaced with the latter. A study of historical patterns in synonymy across eight major insect groups (Gaston and Mound 1993) found that synonymy rates of described species names can vary widely, from a rate as low as 7% (Siphonaptera) to as high as 80% (Papilionidae and Pieridae). A similar study in recently-monographed plant groups found an overall synonymy rate of 66% (Wortley and Scotland 2004), while a later study found a rate of 38% across a subset of all flowering plants (Pimm and Joppa 2015). These numbers include all synonyms ever generated, but in practice scientists are likely to be confused only by synonyms in use recently. Taxonomic redescription is an ongoing process and each redescription may itself be reverted in the future, as philosophies on species boundaries and tools for examining those boundaries change. One study that tallied up invalidation and revalidation rates among North American fossil mammals names between 1850 and 2000 estimated that 24-31% of currently recognized species would eventually be synonymized (Alroy 2002). My previous study of North American birds found that 19% of observed taxonomic changes partially or completed reverted previous changes, suggesting that incorrect proposals are often made even in well-studied, extant taxa (Chapter 2. The tempo and mode of the taxonomic correction process in North American Birds over the last 127 years).

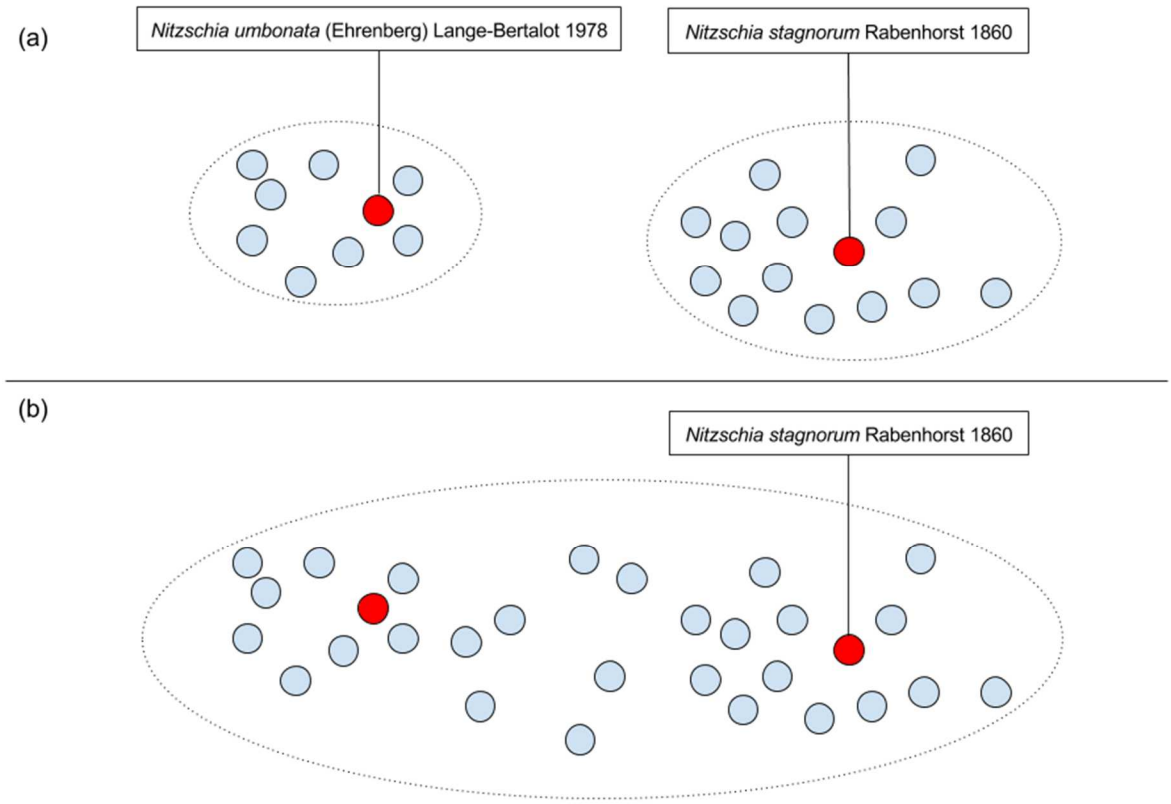


Figure 4. (a) Scientific names at the species level are generally tied to a single type specimen, but have a circumscription that extends around that type to include other conspecific indivudals. (b) As new evidence accumulates, these circumscriptions may be re-evaluated. We can think of the names as being synonymized or as the circumscriptions being "lumped". In either case, the new species retains the same name as the existing species, leading to ambiguity: if a biodiversity record has been identified as Nitzschia stagnorum, does this refer to the original circumscription in (a) or the new circumscription depicted in (b)?

Instead of tracking synonymy, we can also quantify redescription by counting taxonomic changes directly. These changes may occur in the renaming of a taxon from one to another without changing its circumscription (which I refer to as "renames"), through the combination of several taxa into a single taxon ("lumps"), or through the division of a single taxon into several taxa ("splits"). In the case of lumps and splits, one of the resulting taxa retains the same name as one of the source taxa with which it shares a type specimen, but under a different

circumscription (see Figure 4 for an example). In order to compare this with the rate of species being described, we can also count the number of new taxonomic circumscriptions generated by redescription. Tracking circumscriptions distinctly from taxon names was first proposed to allow biodiversity databases to aggregate data from multiple sources that used the same scientific name to refer to different circumscriptions (Berendsohn 1995). However, the rate of circumscription generation can also be used as a measure of taxonomic progress: every new circumscription represents a measurable increase in taxonomic knowledge, even if the new circumscription is itself flawed and will need to be replaced later. The rate at which these circumscriptions are being created, whether by the description of new taxa or the redescription of existing taxa, can be used as a common currency to measure the overall rate of discovery and to partition it among these different processes. It can be tracked at the species level, where it directly affects important measures of biodiversity such as species richness, or at other taxonomic ranks, where it can be used to track how rapidly taxa are being reorganized.

The proportion of shared names and circumscriptions has been used previously as a measure of taxonomic stability: in one study, stability was measured by identifying "reliable" names – those congruent in both name and circumscription – between two primate taxonomic checklists published over a decade apart (Nico M Franz et al. 2016). Similar measures have been used to evaluate the stability of binomial names after major changes to the American Ornithologists' Union's Check-List of North American birds (Rising and Schueler 1972; Olson 1987), but so far

these studies have been limited to comparing one checklist with another. I identified a series of taxonomic checklists of North American freshwater algae species that would allow me to measure the stability of taxonomic checklists over a fifteen-year period. Additionally, as these checklists are used as an index to the United States Geological Survey (USGS) National Water-Quality Assessment (NAWQA) program, established in 1991 to consistently sample freshwater from over fifty major river basins and aquifers across the United States ("The National Water-Quality Assessment Program—Science to Policy and Management" 2010), I could further determine what effect species- and genus-level abundance had on redescription, and what effect redescription had on interpreting biodiversity data in this dataset. These checklists were first created by the Academy of Natural Sciences of Drexel University (ANSP) and maintained by them from 1999 to 2013, and later by the USGS from 2011 to 2017. The most recent version of this checklist ("USGS 12.9"), published by the USGS on May 30, 2017, consists of 11,642 taxonomic units at various levels of resolution, including 3,191 species from twelve classes, dominated by the Bacillariophyceae (diatoms, 59% of species), Chlorophyceae (21% of species) and Myxophyceae (now known as cyanobacteria, 12% of species). I collectively refer to this checklist series as the "Algae Checklists".

While I consider the entire Algae Checklist as a single, continuously updated checklist, the practical significance of understanding taxonomic stability is particularly important for the Bacillariophyceae (diatoms), as diatom occurrence data has been used for water quality assessment globally (Szczepocka et al. 2014; X.

Chen et al. 2016; Belore, Winter, and Duthie 2002), and the records from USGS NAWQA have been used to produce metrics of eutrophication in the United States (Potapova and Charles 2007). Practical uses of diatom occurrence data is hampered by taxonomic uncertainty; for example, several European diatom trophic indexes differ from each other because of the misidentification of some species, lumping of different true species or changing species concepts (Besse-Lototskaya et al. 2011). Determining how stable diatom taxonomy is within this group can provide practical information on how likely these indices are to need to be corrected or refined in the future.

Tracking species names over time also provides information on how the size and composition of genera are changing over time. Within diatoms, several authors have noted that diatom genera tend to be overly broad (Kociolek 1996; Williams and Reid 2006). Kociolek and Williams (2015), noted that ca. 64,000 species of fishes have been classified into ca. 12,000 genera, while a similar number of diatom species have been classified into only ca. 1,200 genera. This trend appears to be reversing in the taxonomic literature: a global catalogue of diatom genera published in 1999 found 907 validly published generic names, of which approx. 20% were described between 1960 and 2000 (Fourtanier and Kociolek 1999). A later study found an additional 93 genera published over just six years (Fourtanier and Kociolek 2003). Using the Algae Checklists, I test whether the number of genera being recognized by this checklist has increased significantly over the last decade and a half.

While checklist series are valuable for measuring how stable freshwater algae taxonomic names and their circumscriptions are, they can also be used to better understand how the different components of the taxonomic discovery process are related to each other. Furthermore, as these names are directly linked to the NAWQA dataset, they can be used to determine if different kinds of circumscription changes are more common for rare or abundant species. To that end, I focused my analysis on the following questions:

1. One of the most striking findings from studying a checklist of North American bird species (Chapter 2. The tempo and mode of the taxonomic correction process in North American Birds over the last 127 years) was a sharply accelerating rate of splitting starting in the 1980s and continuing to the present day. For bird species, I found that this was mainly the result of new discoveries, although it was also partially a result of undoing previous overlumping that had likely occurred as a result of the adoption of the biological species concept. I test whether splitting in North American freshwater algae species is also accelerating, which would suggest that both are the result of current trends in taxonomic practice that apply broadly. If algal species are not accelerating, this would suggest that the pattern of oversplitting ("taxonomic inflation") seen in vertebrate groups might be specific to vertebrate taxonomic communities (Isaac, Mallet, and Mace 2004).
2. Previous studies have shown a sharply increasing number of validly published diatom genera globally between 1980 and 2000 (Fourtanier and Kociolek 1999).

This trend continued into the subsequent decade, with 80 new genera being described in a fifteen year period between 1997 and 2011 (Kociolek and Williams 2015). Furthermore, several diatom genera have extremely large numbers of species – a decade ago, at least ten genera had over 1,000 species, with *Navicula* alone containing over 9,000 species and acting as a 'wastebasket' for any bilaterally symmetrical, raphid diatom that could not be classified into more precisely defined genera (Williams and Reid 2006). If large genera are undergoing the most reorganization, I hypothesize that species-level splits and lumps as well as renames from one genus to another will occur even more often in larger genera than smaller genera when compared with random sampling processes.

3. Since abundant species should be easier to collect, taxonomists may be likelier to initially create overly broad circumscriptions that are later discovered to consist of several distinct species once more evidence accumulates. On the other hand, rare species are harder to collect, leading taxonomists to be more likely to be consider them distinct species that, as more individuals are collected, are discovered to in fact be varieties of a closely-related species. If this is true, I would expect abundant species to be significantly more likely to be split than rare species, while rare species would be significantly more likely to be lumped than abundant species.
4. Despite the possible presence of cryptic species, abundant species are likelier to have been discovered earlier and to have been more thoroughly examined by

taxonomists over time. Since biodiversity data is by definition dominated by its most common species, the effect that taxonomic instability has on the interpretation of most biodiversity data should be minimal. To test whether this was the case, I examined whether the proportion of records associated with more than one circumscription across all diatoms species covered in the list is above a conservative threshold of 5%.

Besides testing these questions, I also calculate a baseline for taxonomic stability within a single checklist over a relatively long period of time. This baseline is not particularly useful on its own outside North American freshwater algae, as this group is likely not representative of other taxonomic groups. However, it does provide an initial set of methods and tools for measurement of overall stability that can be easily extended to other taxonomic checklists and across a variety of taxonomic groups. These methods and tools thus provide needed approaches for a broader and more complete understanding of how all taxonomic discovery proceeds throughout the taxonomic enterprise.

Methods

Lists of recognized taxa. This study is based on three freshwater algae checklists published by the Academy of Natural Sciences of Drexel University (ANSP) between 2002 and 2007, and a later series of forty-five checklists published by the United States Geological Survey (USGS) from 2011 to 2017. I collectively refer to all 48 checklists as the "Algal Checklists". The earliest checklist in this

study, ANSP 2002, was started in October 2002, while the most recent, USGS 12.9, was published on May 30, 2017. Over this fifteen-year period, the number of recognized species increased by 27% from 2,510 species to 3,191 species. The most recent checklist groups these species into twelve classes. These are dominated by Bacillariophyceae (1,891 species or 59.3%), Chlorophyceae (671 species or 21.0%), Myxophyceae (394 species or 12.3%) and Euglenophyceae (151 species or 4.7%), and includes eight other classes (2.6%). Every checklist included higher-level classifications for each species. The list of checklists and the methods used to extract scientific names from them are detailed in an appendix (Appendix 1. Taxonomic checklists used and name extraction methods).

Identifying taxonomic changes. I began by identifying all additions and deletions of names between pairs of taxonomic checklists. Between 2002 and 2017, 1,560 names were added and 2,580 names were deleted, resulting in a net decrease of 1,020 names. I began by excluded changes that involved non-binomial names: those involving higher taxonomic names as well as those involving provisional taxa, whether or not they included a specific epithet as a close match. For example, I excluded "*Aulacoseira* sp. 2 NLS BL cf. *alpigena*", even though it might refer to an individual similar to *Aulacoseira alpigena*. This allowed me to focus exclusively on how species-level circumscriptions changed over time. In all, I found 978 binomial names added and 297 deleted.

I examined the difference between consecutive pairs of checklists individually, and identified cases where a species had been renamed by looking for a

species being added with a similar specific epithet to one that had been deleted, such as the deletion of *Sellaphora eloranta* with the addition of *Sellaphora elorantana* in USGS 12.8 (May 18, 2017). In most cases, species were renamed to a name not previously recognized by the checklist. Where it was previously recognized, or where it was explicitly renamed to a subspecies name, I classified these change as a "lump". Other renames were classified as a "rename", which I used to determining synonymous species names. When a subspecific name (e.g. *Sellaphora pupula* fo. *rostrata*) was renamed to a previously-unrecognized species (e.g. *Sellaphora rostrata*), I categorized this as a "split", as the circumscription of a previous species (in this case, *S. pupula*) was being modified without a change in its name. In some cases, a split occurred where the resulting name was already recognized by the checklist: for example, *Gomphonema acuminatum* var. *brebessonii* was deleted in USGS 5.3 (January 2013), presumably as a split to *G. brebissonii*; however, *G. brebissonii* has been recognized by the Algal Checklists since it was first added in ANSP 2002. I counted these cases as lumps rather than splits: the split must have taken place sometime before 2002, but individuals incorrectly classified as *Gomphonema acuminatum* var. *brebessonii* were now being lumped into the already-recognized *G. brebissonii*. In the case of 30 additions and 7 deletions, I could find no evidence that the name being affected had ever been recognized as a valid species name (e.g. *Phormidium tergestinium*), and so I categorized these changes as "unknown".

I worked primarily by looking for names that shared similar terminal epithets, and checked these changes in nomenclatural guides such as AlgaeBase (Guiry and Guiry 2017) to look for evidence of synonymy. I double-checked these changes by comparing North American Diatom Ecological Database (NADED) IDs across these datasets to identify cases where different names were referred to with the same ID, suggesting that these are synonyms. This method found an additional potential 84 changes, of which I confirmed 46 changes using AlgaeBase. I added these changes to the dataset. Many of these changes occurred in the USGS 6.0 checklist (published Jan 2013), and consisted of species that appeared to be deleted when in fact they were being lumped into other, previously recognized species. In other words, these were cases where the Algal Checklists simultaneously recognized both a species (such as *Scenedesmus bernardii* since 2002) as well as its synonym (*S. acuminatus* var. *bernardii* since 2002), but later deleted one of them, essentially merging the two "taxa" into a single one (in this case, *S. bernardii* in 2013).

Measuring similarity between checklists. To measure the cadence of discovery in similarity-based terms, I measured the Jaccard index of similarity between each checklist and the first checklist. A Jaccard index is defined as the number of shared names in both lists as a percentage of the total number of unique names in both lists, using the following equation:

$$Name\ similarity = \frac{Species\ in\ Checklist\ 1\ \cap Species\ in\ Checklist\ 2}{Species\ in\ Checklist\ 1\ \cup Species\ in\ Checklist\ 2}$$

This provides a simple measure of identity between any two taxonomic checklists, which ranges from 0% (no names in common) to 100% (checklists

containing an identical set of names). Species name similarity is identical to name similarity, but takes synonymy into account, allowing for the same species to be present in the two checklists under different, synonymous names. However, some of those names refer to different circumscriptions in these two checklists as a result of species being lumped or split. I measure this as circumscription similarity, using the equation:

$$Circumscription\ similarity = \frac{Number\ of\ species\ identically\ circumscribed\ in\ both\ lists}{Species\ in\ Checklist\ 1 \cup Species\ in\ Checklist\ 2}$$

This measure will always be less than or equal to the name stability, but will incorporate the effect of circumscription changes. The individual name and stability measurements are plotted visually in Figure 5.

The cadence of taxonomic discovery. I plotted the cadence of discovery in two ways: by cumulatively counting the number of changes that had taken place in this checklist (Figure 6), and by plotting the cumulative number of circumscriptions created by taxonomic description or novel combinations, taxonomic lumps, and splits over time to compare how these processes were progressing (Figure 7**Error! Reference source not found.**). I used a simple linear regression to determine if any of these rates were significantly different from zero. In particular, I test whether the rate of splitting is increasing significantly among North American freshwater algae as it has among North American birds.

Taxonomic discovery is also proceeding in the reorganization of freshwater algal genera. I test whether the number of recognized genera has changed significantly over the course of the Algal Checklists, how the mean size of a genus

has changed over time, and how the proportion of monotypic genera has changed over time.

Using information from USGS NAWQA. I downloaded all 392,757 occurrence records available as "Algae results" in the NAWQA dataset from the USGS BioData website (U.S. Geological Survey 2017a) on July 26, 2017. Each occurrence record has a "Taxon Version Number" of "12.9" for all records, referring to USGS 12.9, the most recent Algal Checklist used in this analysis. These occurrence records are identified to 3,283 taxon names, of which 360,825 occurrence records observed between 1993 and 2014 were identified to 2,557 taxonomic identifications at the species rank or lower. These observations are biased towards diatoms, with 325,535 Bacillariophyceae records (90.2%), but with over ten thousand records for Chlorophyceae (green algae: 17,893 records, 5.0%) and Myxophyceae (blue-green algae: 15,206 records, 4.2%) and over one thousand records for Euglenophyceae (euglenoids: 1,820 records, 0.5%).

The effect of genus size on reorganization. I proposed that the larger genera are most likely to be reorganized taxonomically, including at the species level. However, since larger genera have more species that might potentially be lumped or split, they might be expected to have a larger number of species than other genera for this reason alone. I first tested whether this was true by using a Mann-Whitney *U*-test to determine if genera that contained species that had been lumped or split were likely to be larger than other genera. Then, I used a simple linear regression to determine if larger genera had more lumped or split species. If

this was the case, then the results of the Mann-Whitney *U*-test must be treated skeptically.

Testing whether larger genera were more likely to be reorganized was complicated by the complex trajectories of individual species, that may themselves be split or lumped, or may be entirely moved from one genus to another before being moved back. To simplify this question, I focused on genera that were recognized in the most recent checklist (USGS 12.9, published May 2017). I identified every species that had been previously synonymized in another currently recognized species; for example, *Eulimna minima* is currently recognized as *Navicula minima*, even though *Eulimna* is still recognized as a separate genus. This is concrete evidence that *Eulimna* has been "split from" at the genus level, while *Navicula* has been "lumped into" at the genus level. I used a Mann-Whitney *U*-test to determine if either "split from" or "lumped into" genera were likely to be larger or smaller than other genera.

The effect of abundance on lumping and splitting rates. I hypothesized that abundant species were more likely to be split while rare species were more likely to be lumped. I used the number of observations of each species in USGS NAWQA as a proxy for abundance in North America. I calculated the overall measured abundance of every species in the NAWQA dataset involved in a lump or split, and used the Mann-Whitney *U*-test to determine if species involved in lumps or splits were likely to have more records than others.

The effect of taxonomic uncertainty on interpretation of biodiversity data. To measure whether a significant proportion of biodiversity data may require additional work to interpret because of multiple circumscriptions being associated with the same species name, I determined the proportion of all USGS NAWQA records associated with species with multiple circumscriptions because of lumps and splits. I compared them to an arbitrary threshold of 5%, which corresponds to an average of one in twenty records being associated with scientific names circumscribed in multiple ways.

Results

Changes in the Algal Checklists

The Algae Checklists I studied consisted of three ANSP checklists followed by 45 USGS checklists. The number of recognized species increased from 2,510 species in 2002 to 3,191 species in the most recent checklist studied (May 2017). The first two ANSP checklists have relatively low numbers of species (2,510 and 2,491 species respectively), which increased to 2,988 species in the third ANSP checklist (Supplementary **Error! Reference source not found.**). The first USGS checklist (Feb 2011), which were likely based on a checklist earlier in than the ANSP 2007 checklist, recognized 2,845 species, but this number quickly rose to 3,000 species by August 2012 and to 3,134 species by January 2013. After this point, the number of recognized species remain at or under 3,151 species until increasing to 3,190 species in April 2017 and adding one extra species (*Amphora bicapitata*) in May 2017. I

used a simple linear regression to confirm that this is a statistically significant increase whether we count all checklists ($p < 0.001$, $R^2 = 0.81$), USGS checklists only ($p < 0.001$, $R^2 = 0.73$), or only checklists from January 2013 to May 2017 ($p < 0.001$, $R^2 = 0.61$).

The number of recognized genera has increased from 367 genera in 2002 to a peak of 434 genera in February 2017 before falling to 433 genera in 2017, an 18% increase (Supplementary Figure 8). A simple linear regression shows this as a significantly increasing rate ($p < 0.001$, $R^2 = 0.85$). Most of this increase occurred between 2002 and 2007, when the number of genera increased from 367 to 411; however, the increase since 2011 is still significant ($p < 0.001$, $R^2 = 0.71$). The mean number of binomial names per genus increased significantly from 6.84 to 7.37 ($p < 0.001$). The number of genera containing only a single species in this checklist increased from 156 genera (42.5%) in 2002 to a peak of 183 genera (42.3%) between October 2016 and February 2017 before reaching 182 genera (42.0%) in the last checklist. Some of these are truly monotypic genera, while others contain additional species not found in North America.

Measuring similarity between checklists

Similarity measures between the changes were dominated by three major shifts in the composition of the checklists (Figure 5). Name similarity dropped from 93.8% to 74.7% in the checklists ANSP 2007 (published Sep 2007). Despite a slight reversion towards the first ANSP checklist to 78.7% in the first USGS checklist, the similarity remained steady between 74% and 75% from USGS 1.3 in April 2012 to

USGS 5.2 in Oct 2012, followed by a sharp decline to 66.6% in USGS 5.3 (January 2013). The latest checklist, USGS 12.9 (May 2017), has a similarity of 63.5% with the first checklist. As expected, name similarity is the quickest to change; species similarity – in which synonyms between checklists are considered identical – reaches a minimum of 73.5%. This is a dramatic change: 36.6% of names are not shared between the 2001 and 2017 checklists; a further 10% of names could be matched if we had comprehensive records of synonymy. Over 25% of all the names shared by the Algal Checklists in 2001 and 2017 are unique to one checklist or the other even after taking synonymy into account.

Circumscription similarity closely tracked species similarity, shifting away from it in response to lumps and splits. It reached an overall minimum of 67.6% in April 2017. The difference between species similarity and circumscription similarity

represents the error that could be made if synonymized names are used instead of explicitly defined species circumscriptions.

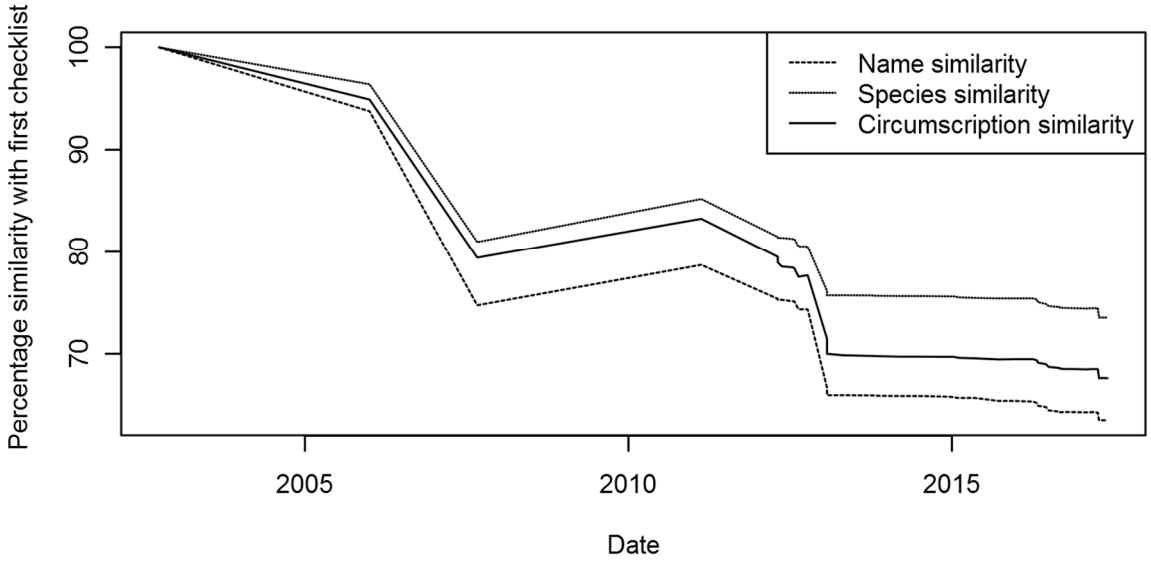


Figure 5. Similarity of each checklist as compared with the earliest one. This includes the name similarity (the number of binomial-level names shared), species similarity (incorporating the effect of synonymy) and circumscription similarity (distinguishing cases where identically named species refer to different circumscriptions).

The cadence of taxonomic discovery

We can obtain a fine-grained measurement of a checklist-specific view of taxon discovery is occurring in this group by identifying the individual taxonomic actions that have taken place over this series of taxonomic checklists. I aggregated species- and subspecies-level changes that affected binomial species composition. This allowed me to ignore subspecific changes that did not affect binomial species composition, such as the deletion or rename of a subspecies that continued to be recognized at the species-level. In all, I identified 1,496 binomial-level changes: 888

additions, 240 deletions, 270 renames (in which one name was replaced by a synonym without any change in circumscription), 69 lumps and 59 splits. I used a cumulative plot to study how quickly these changes are being made over time (Figure 6).

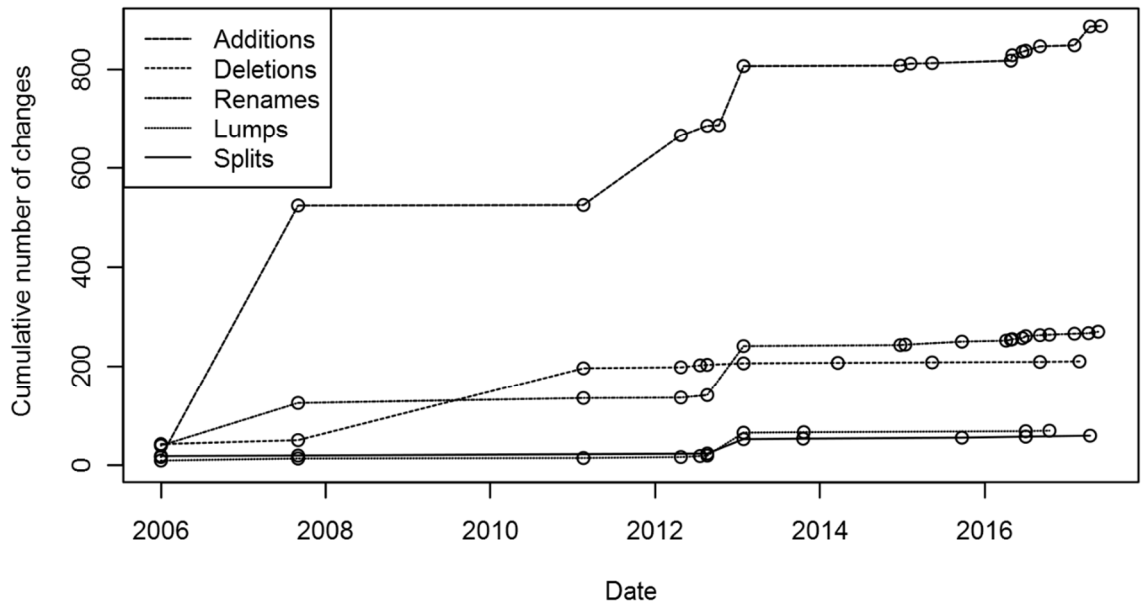


Figure 6. Cumulative numbers of individual changes in the Algal Checklist over time. This graph compares the components of taxonomic change: addition and deletion of taxa, some of which I was able to further classify as renames, lumps and splits.

Additions. I could further break down the 888 additions based on when added combinations were published. I established that 191 additions were published as a species, subspecies or variety in a new description (134) or the publication of a novel combination (60) since 2002, suggesting ongoing taxonomic discovery. The vast majority of the remaining additions were originally described (418) or were described under a new combination (201) before 2002, suggesting that

they represent previous taxonomic discovery, and were added for non-taxonomic reasons, such as because of new evidence that they were found in North America or because they had been accidently overlooked previously. Over the time period of this checklist, we always see more additions than any other change. The number of additions are not changing significantly with time ($p = 0.02$).

Deletions. Of the 210 deletions, most (133) were the result of one set of changes: these names were deleted between ANSP 2007 and USGS 1.0, likely as a result of USGS 1.0 being based on an earlier version of the ANSP checklist. They were then added back in the very next version of USGS in my dataset, USGS 1.3 (versions 1.1 and 1.2 were not published to their website), and so do not represent any additional taxonomic effort. Of the remaining 77 deletions, the majority (46 or 59.7%) were species that were currently recognized, and so I could not draw any clear conclusions as to why they had been deleted from the checklist. They were likely deleted for non-taxonomic reasons, such as the taxon no longer being found in North America. 7 deletions were classified as "unknown" because I could not find any evidence that the names had ever been recognized outside of this checklist, suggesting that they may have been introduced in error. The number of deletions are not changing significantly with time ($p = 0.30$).

Renames. The most widespread taxonomic change I observed were 270 renames. Many of these (96) were the result of a published taxonomic change, of which 55 were published since 2002. 95 renames were minor changes to the spelling of a name, which I classified as typos. The number of renames are decreasing

significantly over time ($p < 0.01$), but this is largely the result of typos: renames excluding typos are not changing over time ($p = 0.17$).

Lumps and splits. Remaining taxonomic changes included 69 lumps and 59 splits. The splits needed no further cleaning, but not all lumps were the result of taxonomic actions. 4 lumps fixed typos in the checklist, such as lumping *Oscillatoria laevis* into *Oscillatoria levis*. As described previously, many lumps appeared to be correcting errors in the checklist where two congruent concepts had been recognized separately, such as when *Jaaginema earlei* was lumped into *Geitlerinema earlei* in USGS 6.0. It was not always clear how to differentiate these cases: when *Gyrosigma parkerii* was lumped into *Gyrosigma wormleyi* in ANSP 2006, should this be considered a true taxonomic lump or a name correction being carried out by lumping? I opted to retain all lumps for my analyses, since even when they do not reflect new circumscriptions being created through taxonomic discovery, they still reflect cases where biodiversity data stored under the two different names, hitherto considered valid, would need to be combined by merging their taxonomic circumscriptions. I summarized the types of lumps and splits in Table 1. Types of 69 lumps and 59 splits in freshwater algae. A complete list is provided in the supplementary materials of this chapter..

Correction type	Type of change	Count	Example
Lump	The combination of multiple species recognized under different, synonymous names.	58	Both *Diadesmis perpusilla* and *Navicula perpusilla* were recognized by the checklist until being combined in January 2013.

	As above, but where the two synonyms appear to be minor spelling corrections ("typos").	5	*Nitzschia liebethruthii* and *Nitzschia liebetruthii* were both recognized before being combined in October 2013.
	A species lowered to a variety	5	*Kirchneriella elongata* being lowered to *Kirchneriella contorta* var. *elongata* in January 2013.
	Two different, synonymous varieties being combined into a single, novel species	1	*Staurosirella pinnata* var. *lancettula* and *Fragilaria pinnata* var. *lancettula* being combined into *Punctastriata lancettula* in January 2013.
Split	Variety or form raised to species	57	*Achnanthes minutissima* var. *jackii* raised to *Achnanthidium jackii* in October 2013
	Variety raised to species, but with a novel name (a "nomina nova" or "nom. nov.").	1	*Navicula ruttneri* var. *capitata* raised to *Sellaphora javanica* in July 2016
	Variety raised to species with a novel name, but it's unclear if this is a novel name or an alternate spelling.	1	*Nitzschia obtusa* var. *kurzii* raised to *Nitzschia kurzeana* in January 2006

Table 1. Types of 69 lumps and 59 splits in freshwater algae. A complete list is provided in the supplementary materials of this chapter.

Thus, I analysed 69 lumps and 59 splits. Neither the number of lumps ($p = 0.98$) nor the number of splits ($p = 0.42$) are increasing significantly over time. While all lumps clearly reflect changes in circumscription, only six of them also involve a change in rank, where species were reduced to varieties or forms. Relatively few checklists contained lumps (only five checklists) or splits (only eight checklists), and only three checklists contained both. A single checklist accounted for most addition through description or novel combinations, lumps, and splits:

USGS 5.3 (January 2013), which can be seen as both a dramatic decrease in similarity (Figure 5) as well as a sharp increase in cumulative additions (Figure 6).

Since both taxonomic descriptions and redescriptions produce new circumscriptions, we can compare these processes directly to each other by comparing the number of circumscriptions produced by each process over time. I divided these processes into four categories: (1) additions through description or novel combinations before 2002 (n = 619), (2) additions through description or novel combinations after 2002 (n = 194), (3) splits modifying an existing circumscription (n = 59), and (4) lumps modifying an existing circumscription (n = 69). Examining this figure (Figure 7) suggests that new descriptions and combinations in the literature prior to 2002 continued to be added until 2013, when they largely ceased, while descriptions and combinations in the literature after 2002 have been added steadily to the checklist and may be increasing sharply in the very recent past. Lumping and splitting contributed new circumscriptions, but in far smaller numbers than additions. This may reflect the small number of checklists that included any lumps (12 out of 48) or splits (8 out of 48). None of these four processes of circumscription generation are increasing significantly over time.

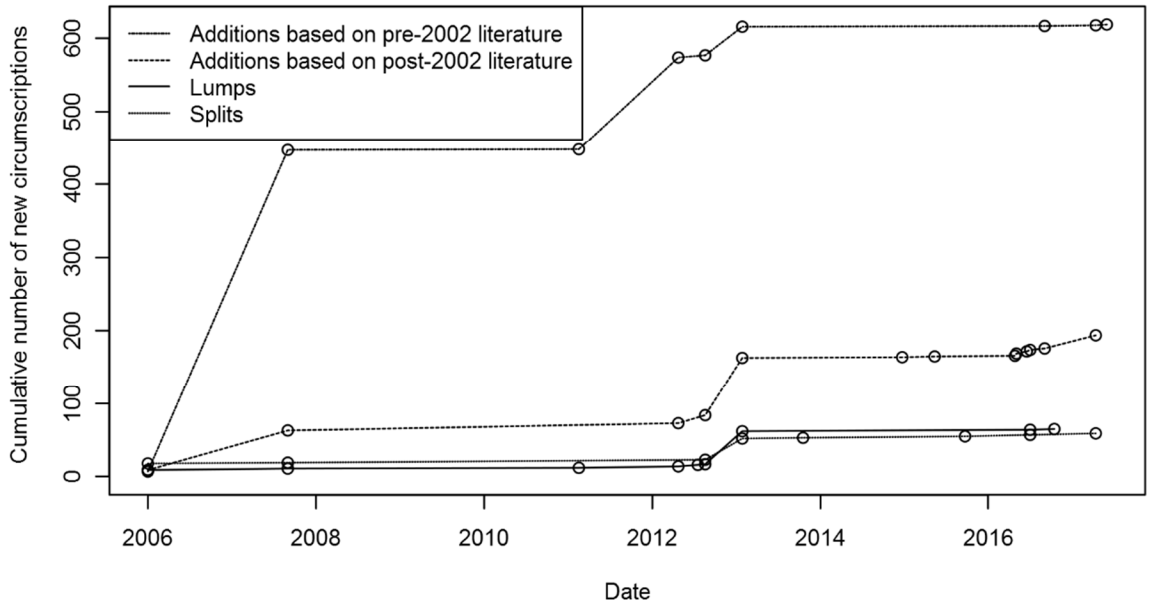


Figure 7. Cumulative number of new circumscriptions added through descriptions and novel combinations, lumps, and splits over time. Three types of taxonomic change create new circumscriptions: additions, lumps, and splits. I was able to further divide additions into those of names described or recombined on the basis of pre-2002 literature, which was published before the first checklist and represents previous description, and names described or recombined on the basis of post-2002 literature, which represents ongoing description.

The effect of genus size on reorganization

I began by determining whether genera containing species that had been split or lumped were larger than genera containing neither. As of USGS 12.9 (May 2017), 433 genera were recognized, which contained a mean of 7.37 species (SE: 0.904). 38 species that were split were contained in 21 genera, which contained a mean of 49.24 species (SE: 12.98). These genera were significantly likelier to be larger than the 412 genera not containing species that were split ($p < 0.001$, Mann-Whitney *U*-test). They were also significantly larger than the 233 genera that contain two or more species in this checklist, suggesting that monotypic genera are

not solely responsible for this pattern (mean: 8.59, SE: 0.84, $p < 0.001$, Mann-Whitney *U*-test).

Lumps can involve multiple species: I found 94 unique names for 86 unique binomial names from 69 lumps. Two genera (*Gloeocystopsis* and *Coscinodiscus*) are no longer recognized, and were excluded. This left 42 source genera, with a mean of 29.57 species per genera (SE: 6.97), which were significantly likelier to be larger than the other 391 genera ($p < 0.001$, Mann-Whitney *U*-test). I repeated this test after excluding one monotypic genus from genera containing species being lumped (*Hannaea*), and found that it was significantly likelier to be larger than the other 210 monotypic genera ($p < 0.001$, Mann-Whitney *U*-test).

One possible explanation for this result is that genera containing more species simply have more opportunities for one of their species to be lumped or split, i.e. this effect might be the result of expected sampling processes in larger or smaller categories. I used a simple linear regression to test whether this was the case. When considering only the genera containing either lumps or splits, I found that the number of lumps ($p < 0.0001$, $R^2 = 61.2\%$) and splits ($p < 0.01$, $R^2 = 37.96\%$) both increased significantly as the number of species increased in the genus.

To look at how genera were being reorganized, I examined the 433 genera recognized in the latest USGS checklist and found 241 names that had synonyms in other currently recognized genera. Taxa (whether species, subspecies or forms) from 35 currently recognized genera had been renamed into 69 currently recognized genera. We therefore have evidence that these 35 genera have lost taxa and these

69 genera have gained taxa through the reorganization of genera. Genera that have gained taxa (mean: 13.68 species, SE: 2.70, median: 7) are significantly likely to be larger than other genera (mean: 6.17, SE: 0.93, median: 2, Mann-Whitney *U*-tests with $p < 0.001$). Genera that have lost species (mean: 31.80 species, SE: 8.53, median: 14) are also significantly likely to be larger than other genera (mean: 5.22, SE: 0.52, median: 2; Mann-Whitney *U*-tests with $p < 0.001$). 18 genera had both gained and lost taxa.

The effect of abundance on lumping and splitting rates

I used the 360,825 occurrence records in USGS NAWQA representing 2,557 species as a proxy for North American abundance of these species. I identified 45 distinct species that had been split in 59 splits (some species had been split multiple times). Of these, 36 species names were observed one or more times in the USGS NAWQA dataset; in the other nine cases, the Algal Checklists had only recognized the variety or subspecies that was raised to a species in the split. These 45 species were associated with a total of 25,167 observations, and were significantly likely to be larger than species that had not been split ($p < 0.001$, Mann-Whitney *U*-test).

I identified 86 binomials that had been lumped in 69 lumps. Only seventeen of these species were still recognized in the USGS NAWQA dataset, as many of them would have been eliminated when being lumped. I tried to improve this resolution by using the taxon name assigned by an analyst during identification (the "Bench Taxon Name") instead of the currently recognized name according to USGS 12.9 (the "BioData Taxon Name"), as this column may store older synonyms

of currently recognized names. This worked, with 53 binomial names matching observations in the USGS NAWQA dataset. Of the remaining 33 binomials, most (24) were eliminated through lumping; 6 were still recognized in USGS 12.9, and so might just not have any associated records in NAWQA; and 3 were only ever recognized as a variety or form, and so were completely eliminated after being lumped into other species-level circumscriptions. The 53 binomials matching bench taxon names in USGS NAWQA matched a total of 13,998 observations, and had a mean of 264.11 observations per species (SE: 116.91). They did not have significantly different numbers of observations when compared with the other 2,532 bench taxon names (mean: 137.00 observations per species, SE: 10.02, $p = 0.94$, Mann-Whitney *U*-test), indicating that species that were lumped were no more abundant than species that were not. Note that this is the only analysis in this chapter that does not use the 360,825 occurrence records in USGS NAWQA representing 2,557 species on the basis of BioData Taxon Names; instead, it uses 360,881 occurrence records representing 2,585 species on the basis of Bench Taxon Names (details of the distinction between these two methods of extracting names are available in Appendix 1. Taxonomic checklists used and name extraction methods).

The effect of taxonomic uncertainty on interpretation of biodiversity data

As before, I used the 360,825 occurrence records in USGS NAWQA representing 2,557 species as a proxy for North American abundance of these species. In 18 cases, the NAWQA dataset contained names that had been shown to

be synonymous elsewhere in the project, such as *Ulnaria ulna* and its junior synonym, *Synedra ulna*. I excluded these from the analysis, resulting in 2,539 species associated with all 360,825 observations. Of this, 188 species (7.4%) had two or more alternate circumscriptions because of lumping or splitting. These species were associated with 70,022 observational records, making up 19.41% of all observations in the USGS NAWQA dataset. The most abundant species with multiple circumscriptions was *Cocconeis placentula*, from which *Cocconeis pseudolineata* was split in ANSP 2006, which had 8542 observational records. One other species (*Achnanthes minutissima*) had over 5,000 observational records, and fifteen other species have over 1,000 observational records. All of these seventeen species were diatoms, except for *Ankistrodesmus falcatus* with 1,393 observational records, which is in class Chlorophyceae. Thus, fewer than 10% of tested species had alternate circumscriptions because of lumps or splits, which could affect the interpretation of almost 20% of all biodiversity records at the species level.

Discussion

Stability of taxonomic checklists

The Algal Checklists have changed substantially over the course of fifteen years, as evidenced by only 63.5% of names, 73.5% of species and 67.6% of taxonomic circumscriptions being shared between the first ANSP checklist (2002) and the most recent USGS checklist (2017). These simple metrics have practical significance for data reconciliation: reconciling data identified using ANSP 2002

against data identified using USGS 2017 should be able to match 64% of shared names without any changes and 74% of names when synonyms are taken into account. Based on the changes in circumscription that we know about in this project, 68% of shared names will match the same circumscription: the other 6% will match a synonymous or identical name with a different circumscription. Thus, over a decade and a half, around a quarter of names can no longer be matched correctly between these two checklists, with a small but significant proportion of all names matching an identical name but referring to a different circumscription.

Given comprehensive synonymy and taxonomic circumscription information, these similarity metrics could be calculated by simply comparing the first checklist to the last rather than comparing across all the lists. However, visualizing the changing similarity over time (Figure 5) across all checklists provides useful information on when and how quickly these changes have taken place. In particular, the Algal Checklists do not simply show steady increases in instability: rather, checklists between 2006 and 2012 show rapid changes that are then reverted, likely representing larger clean-up efforts, while the period from 2013 to 2017 appears to be relatively stable.

Individual taxonomic changes

Dissimilarity within the Algal Checklists is the result of taxonomic changes, primarily additions, renames and deletions, and to a lesser degree lumps and splits (Figure 6). All of these changes appear to be ongoing, with no evidence that any of

them are increasing or decreasing at present. A majority of several types of taxonomic changes were the result of checklist-specific processes, including:

1. Of the 888 additions I examined, 418 additions (47.1%) were of species or subspecies described before 2002, while 201 additions (22.6%) were the result of combinations created before 2002. These are likely the result of species previously described elsewhere in the world being discovered in North America, either because of a range expansion or simply because the species had never previously been detected in North America. These types of additions will be a factor in any regional checklist analyses, and this study suggests that they may form a majority of additions.
2. Of the 210 deletions I examined, 133 deletions (63.3%) were deleted in the transition from ANSP 2007 to USGS 1.0, only to be re-added in the next checklist (USGS 1.3). This is likely a result of the ANSP and USGS checklist series not lining up perfectly, and should be expected whenever different checklist series are synthesized.
3. Of 270 renames, 95 renames (35.2%) fixed minor spelling errors ("typos") in scientific names. Many lumps also occurred when two possibly synonymous species were independently recognized by the checklist, which later needed to be lumped together into a single circumscription.

Taxonomic checklists therefore do not simply reflect the taxonomic changes taking place around them. As with any human endeavour, they may contain errors that persist for some time, and that will eventually need to be corrected through the

use of renames (for spelling errors or changing genera) and lumps (when multiple synonyms of the same taxon are recognized). These are hard to entangle from overall taxonomic processes – for example, a spelling error might be specific to one particular checklist, but it may also have been recognized by several checklists and broadly accepted within the taxonomic community before the error is discovered. In some cases, an incorrect spelling that enters prevailing usage will be preserved to prevent further destabilization of taxonomy, such as through Article 33.3.1 of the International Code of Zoological Nomenclature (Jach 2000). Even where errors are specific to a single checklist, this complicates the task of reconciling checklists with each other or with biodiversity data.

Accumulation of taxonomic circumscriptions. The addition of new taxonomic circumscriptions, whether through "previous description" (the addition of species that were described or placed into a new combination or species before the start of this checklist series), "ongoing description" (the addition of species that have been described or placed into a new combination since the start of this checklist series), or through redescription via lumping or splitting, are continuing at a steady pace (Figure 7), suggesting that this checklist will continue to change in the future, reducing similarity with the first checklist even further.

Circumscriptional changes are particular important in understanding how redescription is occurring and is crucial for data reconciliation, since changes made to the circumscription of a species do not necessary result in a change in taxonomic name. I observed 59 splits and 69 lumps since 2006, the first checklist in which I

could detect a change to a previous checklist. The absolute number of lumps and splits affecting ranks appear larger than the 31 splits and three lumps observed in North American birds between 2006 and 2016 (Chapter 1). However, the number of species in these checklists over that time period are very different, with 826-862 recognized species in the subset of North American birds I studied as compared to 2,491-3,191 recognized species in this checklist. Using the mean of each range, I find 0.037 splits/species in North American birds as compared to 0.021 splits/species in freshwater algae. While 0.004 lumps/species took place in North American birds, the comparable measure is 0.024 lumps/species for all lumps in freshwater algae. The North American bird checklist underwent extensive lumping in the mid-1900s, likely as a result of a shift to the biological species concept, which might lead it to have unusually few lumps in the present.

Of these 69 lumps, however, the vast majority of the lumps we see in this checklist appear to be a by-product of missing synonymy information in the checklist, such as when both *Caloneis molaris* and *Pinnularia molaris* were recognized by the checklist before being lumped in 2013. While such changes clearly reflect changes in circumscriptions, they should not be compared directly with North American birds, where most lumps represent changes in rank. There are only six lumps in which a species is lowered to a variety or form. If we consider only these lumps, there are only 0.002 lumps/species in freshwater algae as compared to 0.004 lumps/species in North American birds. If these numbers are a more accurate measure of taxonomic change occurring outside of these particular checklists, we

observe a potentially shared pattern of splits outnumbering lumps between 2006 and 2016 for both freshwater algae and birds. This pattern is surprising, given that these two groups vary greatly both in biology, in trajectories in species discovery, and in their history of taxonomic practice.

Finally, I found that even a small number of species with multiple circumscriptions – only 7% of the species considered – can affect the interpretation of almost 20% of associated biodiversity data. This suggests that circumscriptional change could have a major effect on the interpretation of biodiversity data, possibly as a result of my finding that species being redescribed are likely to be more abundant than others, discussed in more detail below. Biodiversity databases that do not maintain a checklist with their data, and do not ensure that synonyms are kept up to date and reconciled against each other, will likely see even larger potential errors in taxonomic interpretation. However, the similarity analysis conducted previously shows that 65-68% of names can be matched accurately. Improving this accuracy via technological means, such as through better name matching software (Boyle et al. 2013), should be a priority in biodiversity informatics.

Insights into the process of taxonomy

The hypotheses I tested attempted to determine how different aspects of the taxonomic process were related to each other and to the abundance of different species and genera. To summarize, I found (1) lumping and splitting take place in larger genera, likely because they contain more species than other genera, (2) that

reorganizations between currently recognized genera take place from genera with more species than others to genera that also have more species than others, and (3) that species being split are more abundant than other species, but species being lumped are not. I discuss these in more detail below.

Splits and lumps occur in genera that have more species than other genera, and larger genera have more lumps and splits. This likely suggests that species-level lumping and splitting is not structured by genus, but occurs on a per-species basis – thus, the larger genera simply have more chances to contain a species that is lumped or split than other genera. This suggests that smaller genera containing a few, well-studied species might be more likely to remain stable over longer time periods than larger genera. As the taxonomic correction process continues and fewer lumps or splits remain to be discovered, however, we might expect smaller genera to see more species being redescribed.

When species are moved from one currently recognized genus to another, they tend to move from large genera to other large genera. This suggests that when taxonomists reorganize currently recognized genera – i.e., when they do not merge or split entire genera – they tend to do so by moving species out of large genera into other large genera. Moving species out of large genera suggests that larger genera are being shrunk, exactly as predicted and in line with the significantly increasing mean number of species per genus I observed. Moving them to other large genera suggests that there is a set of small genera that are not being reorganized, either because reorganization of those genera is already complete, or

because larger genera containing more species and subspecies diversity receive more taxonomic attention. As with the previous observation, as larger genera approach completion, we may see a shift to increased attention on smaller genera or, possibly, the completion of genus-level reorganization entirely.

While these two results provide insights into how genera may be undergoing reorganization, they fall short of a complete framework for understanding genus-level description and redescription actions. Given how species-level additions and deletions can be reorganized into taxonomically meaningful actions, such as renames, lumps and splits, it should be possible to reorganize the addition and deletion of species within genera into genus-level lumps and splits, expressed in terms of the species being added to or removed from the genus. This process will be much more complex than the species-level process I describe in this chapter: for example, some species-level actions will have no effect on a genus, such as when a species into split into two species that both remain in the same genus. Such a framework may shed light on how large genera came to be and how they might be reorganized, a question previously investigated in plants (Frodin 2004). A time-based analysis may allow the time span over which such genera were assembled and the rates at which they may be being disassembled to be determined.

Species being lumped and split are more abundant than other species. I initially anticipated that abundance would affect the quality of the initial description: more abundant species would be initially described with an overly broad understanding of its variability, creating an overly large circumscription that

would require later splitting, while less abundant species would be initially described without a complete understanding of its variability, creating an overly narrow circumscription that would later require lumping. I found that splitting does follow this pattern, but lumping does not. Rather than the more sophisticated hypotheses I initially proposed, a simpler explanation for these observations is that there is a certain threshold of collected specimens before taxonomists are confident in their ability to redescribe an already described species. If true, species being lumped or split would be more abundant than species that had been originally described without being redescribed. Given that one in six invertebrate species described at one museum over a decade were described on the basis of a single specimen, and one in four were described from a single locality (Lim, Balke, and Meier 2012), this could impose a significant barrier to the redescription of described species.

More broadly, this could be related to the question of how often taxonomic hypotheses are tested. All analyses implicitly test these hypotheses, but taxonomists may not have the time or resources to thoroughly test each hypothesis. This point was raised by Mann in 2010: "In many cases, of course, species hypotheses will remain untested for a very long time and, especially after passage of the species description into floras and databases, the whole idea of a 'testable hypothesis' begins to seem somewhat esoteric, and we certainly lack the resources to examine every species in detail. Only if the diatom is perceived to be important (e.g. because of unusual abundance, or as an indicator of particular conditions, or

because it is toxic), or occurs somewhere interesting (e.g. in a lake known for endemism in other groups of organisms), is it likely that species hypotheses will be examined critically." (Mann 2010). One of the factors in deciding which taxa are retested might be whether they are known from enough observations for taxonomists to have a clear idea of the variability in the group.

In determining how broadly applicable these results are, it is important to remember that while there was close association between biodiversity data and taxonomic checklists, the two are not in one-to-one correspondence: they were collected over slightly different periods of time (checklists from 2002 to 2017, datasets from 1993 to 2014) and the datasets are biased towards diatom species while the checklists are more evenly split between different algal species. However, they are usually accompanied by a large difference between the groups being compared, suggesting that they will likely continue to differ significantly even if large biases or errors are later discovered. As all of these findings may be specific to taxonomic processes affecting North American freshwater algae, to North American species in general, or to the 21st century. Replicating this study on different groups in different time periods is the only way to distinguish between these different potential causes.

Given the relatively short time period of this study, I did not extensively investigate how often taxonomic changes were undone later in the checklist, as I did with North American birds. However, these "undo" actions do take place in this dataset: for example, *Achnanthes minutissima* var. *jackii* was raised to

Achnanthidium jackii in USGS 5.3, then later lumped back into *Achnanthes minutissima* var. *jackii* in USGS 6.0, before it was finally resplit to *Achnanthidium jackii* in USGS 6.3.

Conclusion

Extracting and analyzing taxonomic changes provides information on synonymy, taxonomic circumscriptions, the cadence of ongoing taxonomic activity, and allows hypotheses about the nature of taxonomic changes to be tested. It is thus valuable for both a theoretical understanding of how checklists change over time, as well as a practical understanding of which taxa have changed the most recently and which might be most likely to change in the future.

Some of the analyses in this chapter make an unusual assumption: that scientists might need to use datasets synthesized from multiple sources, collected using different methods, and identified to taxonomic units by different scientists using different taxonomic views. In practice, scientists will choose their datasets carefully, ensure that the aspects of the data they are interested in have been measured in identical or similar ways, and ensure that their taxonomic units are congruent enough for synthesis to be accurate. Thus, in practice, a 64% match rate between two checklists is unlikely except in very early or especially naïve analyses: the scientists may work to match the remaining concepts or simply exclude them from the analyses if they are not necessary to the hypothesis being tested. However, one of the goals of biodiversity informatics is to reduce this workload for scientists and to facilitate seamless data synthesis (Hardisty and Roberts 2013), so this study

sets out to identify how poorly matching may function in the worst-case with little to no effort put into reconciling it correctly. If developments in biodiversity informatics could increase this from 64% to 94% with the use of online databases of taxonomic circumscriptions and names, digitized checklists for comparison and reconciliation, or through innovations yet undeveloped, this would represent a massive reduction in effort in carrying out large biodiversity data synthesis. The methods used in this chapter might then be useful to track the improvements such technologies bring and to compare their value against matching simply by name, species or taxon concept.

Supplementary Materials

Checklists used in this analysis, a list of all lumps and splits used in the analyses, lists of all binomial changes, other raw data files and R scripts to carry out the analysis are available on Figshare at https://figshare.com/s/0a6832cb6025c53c8a6f. Once this chapter has been published, these data will be made publicly available with its own DOI.

Figure 8. This supplementary figure shows eecognized binomial names, genera, and monotypic genera for each checklist in the Algal Checklist series.

Appendix 1. Taxonomic checklists used and name extraction methods

This study is based on a series of freshwater algae checklists published by the Academy of Natural Sciences of Drexel University (ANSP) between 1999 and 2013, and later by the United States Geological Survey (USGS) from 2011 to 2017. I downloaded ANSP checklists as comma-separated value (CSV) files from http://diatom.ansp.org/nawqa/lists/ansp_taxa_lists.zip and the USGS checklists as Microsoft Excel files from https://my.usgs.gov/confluence/display/biodata/Previous+Versions+of+the+BioData+Algal+Taxonomy. I included both sets of checklists in my supplementary materials for this chapter (https://figshare.com/s/0a6832cb6025c53c8a6f). The most recent

version of this checklist consists of 11,642 taxa, identified to various taxonomic classifications or, in many cases, classified into general groupings such as "Undetermined Rhodophyte". In this appendix, I explain which checklists I included and excluded from the analysis, and how I identified the species-level names associated with each taxonomic name.

I included three of the six ANSP checklists. Two ANSP Checklists ("NAWQA 1994-Starts" and "NAWQA 1997-Starts") were ignored because they were not available for download from the ANSP website and lacked higher taxonomic data. Furthermore, NAWQA 1994-Starts is a limited list, with only 1,357 taxon names as compared to 5,089 taxon names in NAWQA 1997-Starts.

Three ANSP Checklists with higher taxonomy were available for download from the ANSP website: "NAWQA 2001-Starts" (October 2002), "NAWQA 2004-Starts" (January 2006) and "ANSP 2007" (September 2007). All of these checklists were updated after initial publication until being replaced by the next checklist, and so do not necessarily represent views at a single point in time. Because of this, the final ANSP list (ANSP 2007) contains some changes undone by the first USGS list (USGS 1.0), which I attribute to different curation practices in these two lists. Between 2002 and 2007, the number of taxon names (including higher taxonomic names as well as genus-only names such "Nitzschia *sp.*") in these checklists increased from 5,131 to 6,506 in this period, while the number of binomial names increased from 2,510 to 2,988.

In order to link the ANSP series with the USGS series of checklists, I tried to identify the version of ANSP that USGS 1.0 was based on. ANSP 2011, the final ANSP checklist, had significant discrepancies with both the checklist before it (ANSP 2007) as well as the one after it (USGS 1.0). At the binomial name level, this discrepancy translated to 283 binomial names added and 841 deleted between ANSP 2007 and ANSP 2011, and 781 binomial names added and 363 deleted between ANSP 2011 and USGS 1.0. By contrast, ANSP 2007 to USGS 1.0 differed in only 11 additions and 154 deletions. Therefore, I decided to exclude ANSP 2011 from my analyses. Even so, many changes made between ANSP 2007 and USGS 1.0 were reverted in USGS 1.3, suggesting that USGS 1.0 was in fact based on an earlier version of ANSP 2007.

The USGS has published 12 major and 33 minor checklists of this data. The first, USGS 1.0, was published on February 17, 2011, while the most recent is USGS 12.9, published on May 30, 2017. Over this time period, the number of all taxonomic names decreased from 6,020 to 4,111 names, with a sharp drop in names at USGS 2.0 (July 18, 2012) as many names identified only to genus were deleted. The number of recognized binomial names varied from 2,845 to 3,191 recognized names.

These checklists included higher taxonomic names, species, varieties as well as group names ("Undetermined Rhodophyte"). Scientific names were encoded using two main strategies:

1. NAWQA 2001-Starts, NAWQA 2004-Starts and ANSP 2007 checklists had different components of the name encoded in "genus", "species", "variety" and "form" columns, with "second_taxon_name", "third_taxon_name" and "fourth_taxon_name" columns in between them. I concatenated these columns together and filtered out obviously incorrect scientific names, such as those starting with "Undetermined …" or "Unknown …".
2. USGS Checklists used a three-level name-determination process (https://my.usgs.gov/confluence/display/biodata/BioData+Taxonomic+System), in which names are initially entered as bench-identified taxa, then potentially synonymized to BioData taxon names, and finally linked to published taxonomic names. The bench-identified taxa are never deleted, but may be deprecated or retired. Published taxonomic names link names with other checklists such as ITIS (U.S. Geological Survey 2017b); in a very few cases where an exact match could not be found, the published name may refer to a genus rather than a species: for example, *Scenedesmus obscura* in USGS 12.9 simply uses a published name of "*Scenedesmus*", ignoring the specific epithet entirely. Because of these shortcomings and since BioData taxon names are intended to provide a standard nomenclatural system for this taxonomic system, I used them in identifying changes in the sets of recognized names.

Scientific names were extracted from every row of every checklist, except for several rows in USGS 1.0, which had genus-level bench taxon names but with blank

BioData taxon names. Since I am primarily interested in species level names, I ignored these genus-level names.

CHAPTER 4. SCINAMES: A TOOL FOR ASSEMBLING DATASETS OF TAXONOMIC CHANGES TO IMPROVE BIODIVERSITY DATA RECONCILIATION USING TAXON CONCEPTS

Introduction

Aggregating biological data from multiple sources allows research on larger scales and over longer time periods than would be possible with a single dataset. Digitally accessible information about species distributions (*sensu* Meyer et al. 2015) is currently being produced at an unprecedented rate from a wide variety of sources, including digitized museum collections (Constable et al. 2010), individual observations from amateur naturalists (Barve 2014) and historical field notebooks (Thomer et al. 2012). These observations result in useful summary products such as coarse-scale range maps (Jetz, McPherson, and Guralnick 2012), and may be augmented by other forms of media, such as photographs of a field observation, records of ecological interactions, tracking information from radio transmitters or detailed phenotypic measurements (Peterson et al. 2010). A drive towards standardization around the Darwin Core format (Wieczorek et al. 2012) has allowed these data to be pooled together, first into geographically or taxonomically focused databases such as iDigBio or VertNet, and from there into global databases such as the Global Biodiversity Information Facility (GBIF). While digitized biological observation can now be accessed and downloaded in an interoperable format from any of these sources, a crucial piece of the global biodiversity infrastructure is still incomplete: aggregating these data on the basis of taxonomy remains extremely challenging because of the ambiguity of scientific names (Remsen 2016).

The identification of an individual that is observed or collected is generally made by using a field guide, taxonomic checklist or another taxonomic authority and recorded as a scientific name. However, the same scientific name may be interpreted by different taxonomic authorities to refer to different meanings of the set of individuals it includes (its circumscription or "taxon concept" *sensu* Nico M Franz et al. 2008). Where the taxonomic authority originally used to make the species determination for a particular dataset is known, the circumscription intended can be determined easily. But aggregating data between two datasets requires knowing they those circumscriptions relate to each other. Is a citizen scientist observing *Branta canadensis* in Alaska in 2017 using the same species circumscription as the *B. canadensis* collected for a museum in New York in 1893?

Here, I describe a software tool that allows scientific names in multiple taxonomic checklists to be cross-mapped to each other and to names in biodiversity datasets, allowing biological data to be reconciled by name, by synonyms, or by taxon concept. I also present a data format for sharing this cross-mapping information as a biodiversity dataset, allowing taxonomic hypotheses and changing taxonomic knowledge to be shared and reused for future analyses. Before introducing the tool more fully, I first discuss in more detail the challenges in using scientific names.

Challenges in using scientific names

Using scientific names to refer to biological taxa presents three main challenges, listed here in increasing complexity:

1. Binomial names consist of a genus name and a specific epithet. When a species is moved from one genus to another, its genus name is changed, and its specific epithet may also be corrected to maintain grammatical unity (Ride et al. 1999,

art. 31.2; McNeill et al. 2012, art. 23.5). Two different databases may use such synonyms to refer to the same species. For example, the tiger was named *Felis tigris* Linnaeus 1758 before being moved to the genus *Panthera* (Pocock 1929). It is currently recognized as *Panthera tigris* (Linnaeus 1758). This challenge only affects species and infraspecific names.

2. Taxonomic determination uses the oldest validly published name associated with the taxon of interest. Therefore, newly discovered literature might affect which name is considered the oldest. For example, different taxonomists may have described the same species using different individuals, as happened for the platypus (Grant 1989). Ambiguities in old descriptions, particularly those not based on preserved specimens or where these specimens have been lost, will need to be resolved by later taxonomists. An example described in detail by Olson and Reveal (2009) is the Blue-winged Warbler, originally named *Certhia pinus* by Linnaeus, based on two illustrations and a description that were later discovered to be based on two different species. Wilson, acting as first reviewer, renamed this warbler *Sylvia solitaria* in 1810. As another species had been previously described under that name in 1808, it was left to Olson and Reveal to provide a unique name for this species: *Vermivora cyanoptera* Olson and Reveal, 2009.

3. Taxonomic names circumscribe a set of individuals that are included within the taxon. This circumscription may be subsequently changed by taxonomists in light of new evidence. At higher taxonomic levels, these can be described by enumerating the taxa included within it: for example, a family may be described by a list of the genera it contains. Subspecific groups may not be present in all species, however, so not all species can have its contents precisely enumerated in this way. Such alternate circumscriptions cannot easily be disambiguated: under

current rules of nomenclature (Jach 2000; McNeill et al. 2012), when one putative species is determined to contain multiple species ("split") or when multiple putative species are determined to be a single species ("lumped"), one of the resulting species retains the same name as the original species. Thus, a single species name can accumulate multiple alternate meanings associated with it over time, and a published instance of a name might be intended to refer to any of them. Species names can be disambiguated by including a citation to a publication in which the species name is clearly defined (Berendsohn 1995; J. B. Kennedy, Kukla, and Paterson 2005; N. Franz, Peet, and Weakley 2008).

Reconciliation of biodiversity data necessitates identifying and accounting for these differences so that data can be aggregated using a single, consistent view of taxonomy, such as that presented in a single taxonomic checklist. In response to this need, taxonomic checklists are increasingly becoming an important part of biodiversity databases. Some databases, such as GBIF, iNaturalist, NCBI GenBank, Tropicos and AmphibiaWeb, have developed in-house taxonomies that serve both to ensure that scientific names are spelled consistently and to provide hierarchical navigational tools. Some biodiversity databases are primarily taxonomic resources, such as ITIS, the Catalogue of Life, the Plant List, the Reptile Database, and Amphibian Species of the World. Each of these checklists are continually updated as species hypotheses are tested and potentially falsified on the basis of new evidence (Sluys 2013). Thus, aggregating data requires not just reconciling all data sources against a single checklist, but against multiple versions of that checklist and possibly against different checklists used by different data sources. This process could be greatly simplified if different versions of different checklists had previously been cross-mapped with each other, providing an initial dataset that further cross-mapping could be built upon.

An alternate approach to building cross-mappings of taxonomic hypotheses is based on identifying and transmitting relationships between taxon concepts using a formalized schema (Taxonomic Names and Concepts Interest Group 2006; N.M. Franz and Peet 2009). A few databases of taxon concepts have already been built, based on taxonomic revisionary work (Weakley 2015) and on a customized solution for integrating myriad bird checklists (Lepage 2017; Lepage, Vaidya, and Guralnick 2014). An ecosystem of taxon concept-based tools is also being created: upstream tools to generate new taxon concepts, such as by extracting them from species descriptions (Cui et al. 2016), as well as downstream tools that use taxon concepts to generate new biological insights, such as logical reasoners that can clarify and explicate taxonomic reasoning (M. Chen et al. 2014). These developments have led to studies that explicitly consider taxon concepts in ecological analyses (Faria et al. 2013; Peet, Lee, and Boyle 2012). These datasets provide a valuable source of taxonomic cross-mapping information, but there is as yet no way to aggregate data using taxon concept relationships alone.

In this paper, I describe an open-source Java application that can be used to generate such cross-mapping data from a series of taxonomic checklists and to reconcile biodiversity data using these cross-mappings. While the ability to help users quickly manage taxonomic content from updating checklists is the primary motivation for this application, it can also quickly provide a means to quantify and visualize name changes, from simple ones such as additions and deletions, to more challenging ones, such as taxonomic concept change. My application provides a graphical user interface that allows the cross-mapping to be displayed and edited, allowing potential errors to be quickly identified and corrected. It also allows cross-mappings to be treated as a biodiversity data resource independent of the data being aggregated, allowing it to be published and reused. Finally, I have included

searching, testing and bulk editing features I found useful when working with scientific names in taxonomic checklists and biological datasets. I also define an XML data schema for storing and publishing these cross-mappings.

Design goals

My primary design goal for SciNames was to provide an interface for datasets of taxonomic cross-mappings to be built, shared and applied to biodiversity data. For this initial version, I focused on species names for analyses, but the data format and software should be able to store information at any taxonomic rank. In designing this software, I identified four main goals:

1. **Identifying cross-mappings between a series of taxonomic checklists.** This goal consists of identifying the additions and deletion of scientific names that have taken place between two or more taxonomic checklists or over a series of taxonomic checklists, and allowing users to annotate some changes as the result of a species being split into several species or lumped into a single species. This functionality is similar to the software TaxoNote Comparator (Morse et al. 2003) and Taxonaut (Ytow 2016), but those focus on comparing names and hierarchies rather than identifying and sharing cross-mappings.
2. **Allow taxon concepts based on these cross-mappings to be used to reconcile data.** This goal consists of using the cross-mappings generated in the previous goal to identify taxon concepts, and then to aggregate data within each taxon concept. SciNames supports any tabular data that can be represented as a UTF-8 encoded comma-separated values (CSV) file, and supports export of the data in the same format.

3. **Store and share cross-mappings in a file.** I chose a native representation in XML to provide a human-readable and editable file format for publication. I formalize version 1.0 of this representation as an XML schema using XSD, allowing it to be validated by XML validation tools. This allows the cross-mapping information to be shared, built upon and reused.
4. **Validate and summarize cross-mappings.** Taxon concepts are complex types that can be challenging to validate by eye. I have therefore included validation tools that detect common errors, such as an explicit "deletion" of a name that was not previously recognized by a particular checklist, and allow them to be corrected quickly. I also include tools that infer potential taxonomic changes automatically, providing recommendations that may speed up the reconciliation process.

Data model and XML representation

My data model is based on three main data types: projects, datasets and changes.

A **project** consists of a set of datasets arranged in a particular sequence, usually in chronological order of publication. It may be used to represent different taxonomic checklists reconciled against each other or a single taxonomic checklist changing over time. Projects can be saved, loaded and shared as an XML file, which can be validated by the provided XML schema file (https://github.com/gaurav/scinames/blob/master/xsd/scinames.xsd).

A **dataset** consists of tabular data, with each row associated with one or more scientific name, as well as a list of taxonomic changes that took place at this dataset. This can be used to represent a variety of biodiversity data, from a simple

list of species names to complex data serialized as JSON (Bray and Google Inc. 2014). Datasets may have a publication date. A dataset may further be treated as a **taxonomic checklist**, in which case it is assumed to be a comprehensive list of every species recognized at that place and time. Every dataset consists of a number of **columns**, **rows**, and a list of **changes** that took place in this dataset. Rows are stored as key-value pairs. Values must be encodable as UTF-8 strings, but complex data can be stored in a single key using JSON or another string-based representation.

A **change** represents an individual taxonomic change made in a particular dataset. Each change can be of one of five change types: "addition", "deletion", "lump", "split" and "rename". Each consists of a set of names being added (the "to" names) and a set of names being deleted (the "from" names). Some types are constrained in the number of from and to names they may contains: additions and deletions may only have one, while renames must be from one name to one other.

SciNames differentiates between **explicit changes** and **implicit changes**. Explicit changes denote changes that explicitly took place in a particular dataset. In SciNames, they are added by users or generated by built-in tools. They are checked by validation tools and may be edited or deleted if they are found to be incorrect. Implicit changes are generated for a checklist when a project is loaded into SciNames or when new checklists are added or rearranged. They consist of all the changes necessary to bring the names from the checklist immediately preceding the present checklist in line with the names in the present checklist. For example, if checklist B has five more names than checklist A, SciNames will generate five "addition" changes for these names. Note that implicit changes are generated solely for checklists, but not for any other datasets. A dataset may therefore contain

information on a small number of species within a larger taxonomic checklist without affecting the list of species recognized at that point in time.

SciNames also tracks two types of data that are not stored as part of a project, but are calculated whenever a project is loaded or updated: **name clusters** and **taxon concepts**. Name clusters are sets of synonymous names that are identified using "rename" changes — when a species is renamed from one name to another, they are treated as synonymous everywhere in the project. Name clusters are available for all scientific names provided in the file (see *Text processing of scientific names* below), but most analyses are run on **species name clusters** that only use binomial names, ignoring both genus-level names and subspecific differences. A name cluster represents a "nominal concept" – a taxon concept consisting of a name with no additional information on how that name is intended to be used (Taxonomic Names and Concepts Interest Group 2006). Every taxon concept with a name is therefore a subset of its name cluster. Taxon concepts refer to a particular circumscription of a particular name. These are identified using "lump" and "split" changes: when a name emerges from a lump or a split, it is assumed to have a different circumscription than it did before the change.

A complete description of the XML format is provided in its XSD representation on GitHub at https://github.com/gaurav/scinames/blob/master/xsd/scinames.xsd.

Text processing of scientific names

Scientific names are particularly difficult to process in computer software: they are generally short name-strings that can be represented in many similar textual representations (compare "*Branta canadensis*", "*Br. canadensis*" and "*B. canadensis*") or divided into multiple columns – the Darwin Core standard allows

separate columns for *dwc:genus* and *dwc:specificEpithet*, but some datasets combine them (along with authority and subspecific information) in the *dwc:scientificName* column. They may contain infraspecific epithets, such as a subspecies name or a variety or form, which may be spelled out (e.g. "*Panthera tigris* ssp. *jacksoni*") or represented as a trinomial (e.g. "*Panthera tigris jacksoni*"). Certain terminology can be used to indicate a genus-level identification (e.g. "*Panthera* sp."), multiple species in a single genus (e.g. "*Panthera* spp.") or a species that is similar to a known species ("*Panthera* aff. *tigris*" or "*Panthera* cf. *tigris*"). Any scientific name may also contain an authority: a short citation to the original description of the scientific name. Specialized software to parse a scientific name is available (Mozzherin, Myltsev, and Patterson 2017), but components of scientific names are divided into separate fields in many biodiversity datasets.

Our model of a scientific name consists of two required elements: a **genus name** and a **specific epithet**. If the epithet is absent, this indicates that the name refers to a part of a genus but has not been identified to a single species. A scientific name may also have an **infraspecific name**: this includes all the information provided in the scientific name that is not part of the genus name or specific epithet. SciNames treats this as a single, unparsed string that opaquely identifies a part of a species (that is itself identified by the genus and specific epithet). This broad definition of infraspecific epithets includes subspecies, forms, varieties, population-specific identifiers, authority information and references to literature.

When importing a dataset from a file, SciNames provide an extensible system of name extractors to extract scientific names spanning multiple columns. Built-in extractors are listed in Table 2. These can be chained together using the "or" binary operator, which is always evaluated from left to right. Thus, the extractor "genusAndEpithets(genus, species) or scientificName(species)" will attempt to

extract a genus name from the column 'genus' and a specific epithet from the column 'species'; if either column is blank or empty, the SciNames will next attempt to extract an entire scientific name from the column 'species'. By default, loading a dataset will use a preconfigured sequence of extractors that match many common dataset patterns. Once a file has been loaded, the name extractors used can be changed on the fly.

Name extractor	Function	Example
scientificName(column)	Extracts genus, specific epithet and any subspecific epithets from a column	"Panthera tigris jacksoni Luo et al., 2004" would be parsed as genus (*Panthera*), specific epithet (*tigris*) and infraspecific information (*jacksoni* Luo et al., 2004).
binomialName(column)	Extracts a scientific name from the column, then truncates it to the binomial name.	"Panthera tigris jacksoni Luo et al., 2004" would be parsed as genus (*Panthera*) and specific epithet (*tigris*), with other infraspecific information discarded.
genusAndEpithets(genus, specificEpithet)	Extracts genus and specific epithet from two different columns.	Given "Panthera" in a column "genus" and "tigris" in a column "specificEpithet", a name with genus (*Panthera*) and specific epithet (*tigris*) will be extracted.

genusAndEpithets(genus, specificEpithet, infraspecificInformation)	Extracts genus, specific epithet and subspecific information from three different columns.	Given “Panthera” in a column “genus”, “tigris” in a column “specificEpithet” and “jacksoni” in a column “infraspecificInformation”, a name with genus (*Panthera*), specific epithet (*tigris*) and infraspecific information (*jacksoni*) will be extracted.

Table 2. Name extractors provided in SciNames

While currently specific to SciNames, this system of combining information from several different columns or extracting particular information from a column may be useful in other biodiversity tools.

Name reconciliation and data aggregation

When a project is loaded into SciNames, it builds *name clusters* of synonymous names. These are sets of synonymous names based on “rename” changes: if one name is renamed to another, they are both considered synonyms with identical taxon concepts throughout the project. When datasets are first added to a project, no “rename” changes are present, and so no synonyms are known. These can be added manually or inferred using the “Infer Changes” tool (Appendix 1: Inferring changes).

Data aggregation can be carried out using scientific names, name clusters or taxon concepts using the “reconcile data” command. Taxon concepts are created by dividing name clusters at “lump” and “split” changes; without lumps and splits in the project, differentiation on the basis of taxon concepts is not possible. Internally,

this is a two-step process: first, scientific names, name clusters or taxon concepts in the datasets of interest are reconciled against each other. Data is then aggregated on the basis of these reconciled concepts.

Users need to make two main decisions when aggregating:

1. They need to choose the dataset whose names, name clusters or taxon concepts are used for aggregation. Often, this will be a single dataset that the user wants to supplement with data from other datasets. SciNames can also aggregate data on the basis of every name, cluster or concept used anywhere in the project.
2. Secondly, they need to choose a dataset containing the data they want to aggregate. Alternatively, SciNames can aggregate data from every dataset in a project.

Aggregation produces a single dataset containing a list of names or taxon concepts from the name dataset with data from the chosen dataset. These can be exported as comma-separated value (CSV) files for downstream analysis in other software such as R.

Visualization

Taxonomic change and name stability are complex processes that are difficult to represent visually. The clearest descriptions of these processes can be obtained by measuring stability as a percentage of recognized names or name clusters (see Chapter 3) or by measuring how often taxonomic changes take place and the number of taxon concepts per recognized name (see Chapters 2 and 3). Other possible metrics include changes in the number of species per genus over time, in numbers of added or deleted species, or in the total number of explicit and implicit changes made (see Case Studies below). In many cases, these values graphed over

time provide a visual guide to changing rates over time, allowing abrupt changes in stability, long-running trends in particular types of changes or in the number of species or genera recognized to be detected.

I developed a visualization based on the stability between checklists calculation from Chapter 3. We can measure the stability between two checklists using the Jaccard index or "Intersection over Union" metric:

$$Similarity(A, B) = \frac{Names\ in\ checklist\ A\ \cap\ Names\ in\ checklist\ B}{Names\ in\ checklist\ A\ \cup Names\ in\ checklist\ B}$$

We can further distinguish between name stability, in which two names are considered identical if they are exactly identical, and name cluster stability, in which two names are considered identical if they are synonymous with each other. This allows us to visualize how name and name concept stability changes over time by comparing each checklist with the first checklist in our project.

Expected workflow

While SciNames is designed to be flexible in terms of inputs and processing steps, it is likely that most users will follow an initial sequence of steps to ensure that datasets loaded by them have been loaded correctly, and to maximize the opportunity to back-track from an incorrect analysis. The workflow I used to prepare the test cases followed this sequence:

1. **Prepare datasets for import:** Taxonomic checklists must be converted into UTF-8 encoded, comma-separated value (CSV) files before import to SciNames. Both Microsoft Excel and Apple Numbers can produce such files, although Excel will default to UTF-16 files unless the "CSV UTF-8" option is explicit selected at export. Each row in the file can contain one or more names, either in a single column or in separate genus and specific epithet columns. All data must be

encoded as UTF-8 strings without control characters. SciNames can load Microsoft Excel files directly, but requires far more memory to load these files than CSV files. Once they have been saved as XML files, no additional memory is needed.

2. **Import data and set date:** CSV files can be added to a project by dragging them into the main window or by clicking the "Add Dataset" button. They can be renamed and a date of publication set. Note that approximate dates are allowed, such as "August 2016" (which is treated as "August 1, 2016") or "2011" (which is treated as "January 1, 2011"). Checklists do not need to be dated, but may be useful for measuring rates of taxonomic change.
3. **Check name import:** While SciNames attempts to guess how names are being stored in a checklist, it may not guess correctly. The number of recognized names is displayed in the main window (Figure 9), but datasets can be double-clicked to display a list of data rows and the names extracted from them (Figure 10). The name extractor may need to be modified to customize it to a particular dataset (see Text processing of scientific names above).
4. **Validate:** The validation tools can be used to validate the dataset to check for rows from which names could not be extracted, scientific names that contain non-ASCII characters, and other tests to ensure that data has been imported correctly. This should be carried out before using an imported project for analyses.

When reconciling datasets with taxonomic checklists, I recommend the following order:

1. Load taxonomic checklists using the steps described above. I recommend loading a single checklist, setting up the name extractors correctly, and then loading the

remaining checklists using the correct name extractor. This project should be saved separately in case later changes need to be undone.

2. Taxonomic changes can be inferred from the raw data using various techniques. In particular, synonymy and taxonomic changes (see Case Study 1: AmphibiaWeb) may be improved by inferring new changes.

3. The project at this point contains cross-mappings between a series of taxonomic checklists. This can be published, ideally in a source code repository like GitHub that would allow it to be corrected and improved by multiple users. Analyses of checklist stability and changes over time can be visualized at this point.

4. For dataset reconciliation, one or more datasets will need to be added to the project. The "Reconcile Data" tool can then be used to match the data with scientific names.

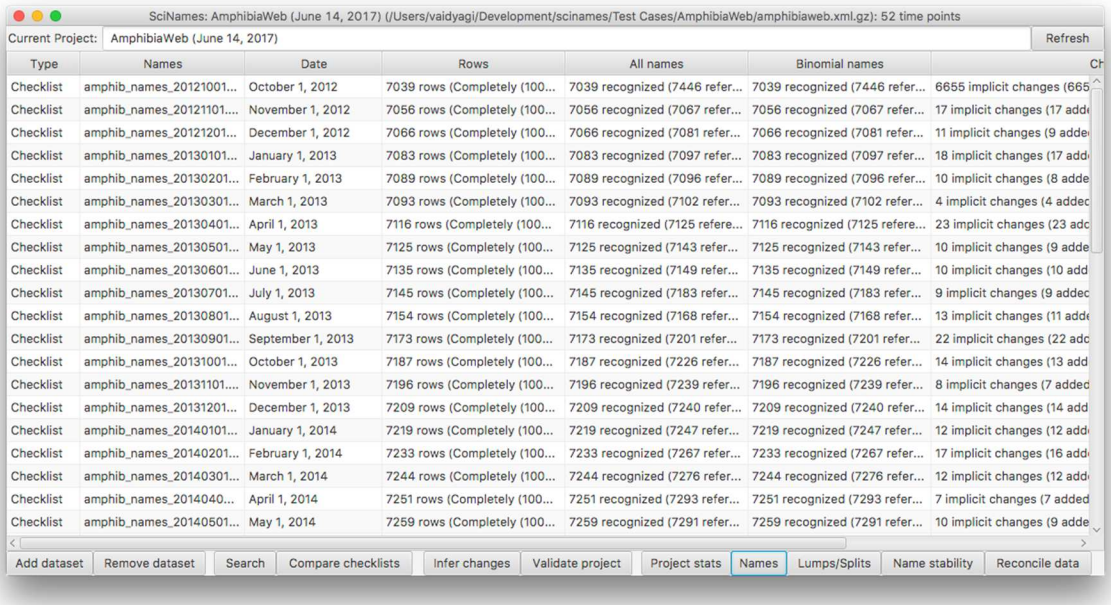


Figure 9. A project containing 52 checklists in SciNames.

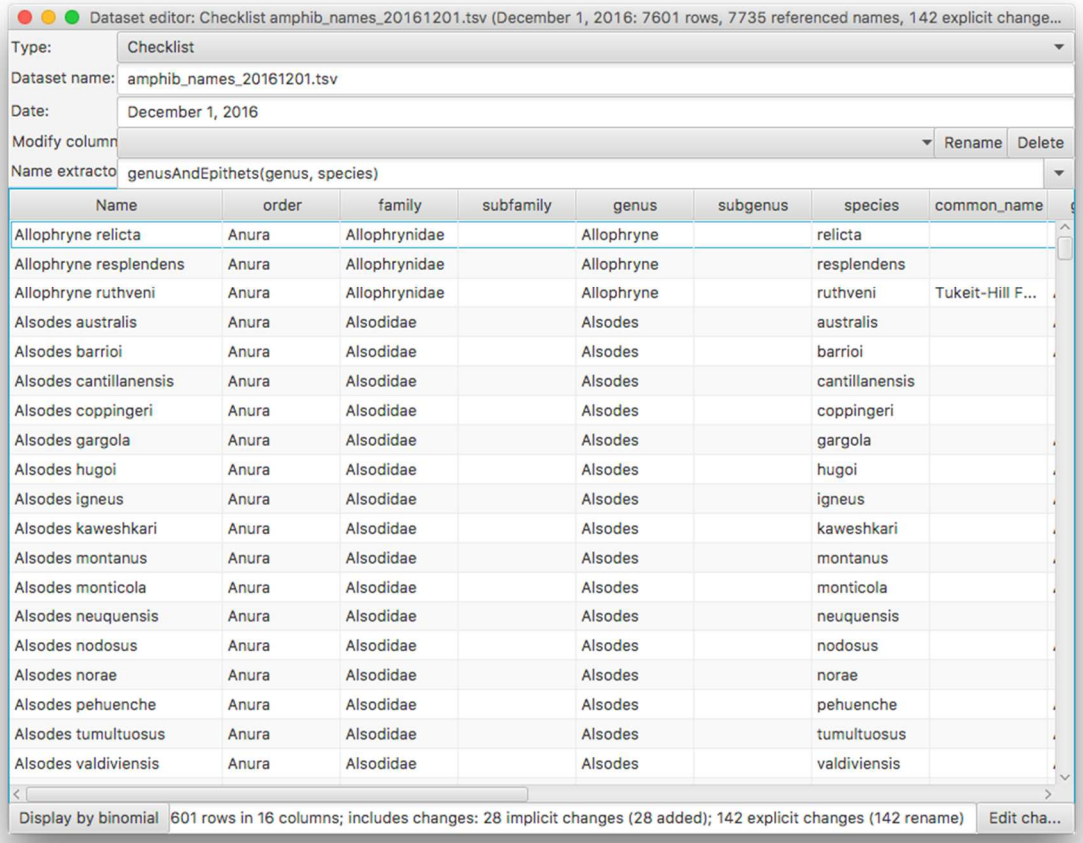
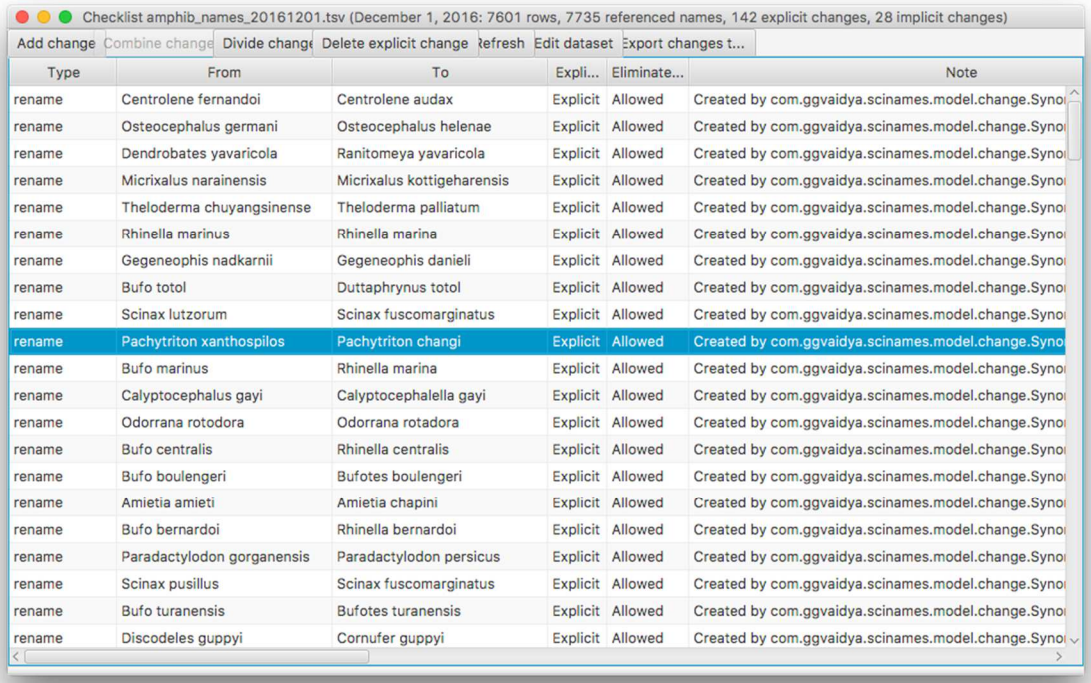


Figure 10. Dataset editor for a single checklist or dataset in SciNames.



Type	From	To	Expli...	Eliminate...	Note
rename	Centrolene fernandoi	Centrolene audax	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Osteocephalus germani	Osteocephalus helenae	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Dendrobates yavaricola	Ranitomeya yavaricola	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Micrixalus narainensis	Micrixalus kottigeharensis	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Theloderma chuyangsinense	Theloderma palliatum	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Rhinella marinus	Rhinella marina	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Gegeneophis nadkarnii	Gegeneophis danieli	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Bufo totol	Duttaphrynus totol	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Scinax lutzorum	Scinax fuscomarginatus	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Pachytriton xanthospilos	Pachytriton changi	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Bufo marinus	Rhinella marina	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Calyptocephalus gayi	Calyptocephalella gayi	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Odorrana rotodora	Odorrana rotadora	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Bufo centralis	Rhinella centralis	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Bufo boulengeri	Bufotes boulengeri	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Amietia amieti	Amietia chapini	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Bufo bernardoi	Rhinella bernardoi	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Paradactylodon gorganensis	Paradactylodon persicus	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Scinax pusillus	Scinax fuscomarginatus	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Bufo turanensis	Bufotes turanensis	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno
rename	Discodeles guppyi	Cornufer guppyi	Explicit	Allowed	Created by com.ggvaidya.scinames.model.change.Syno

Figure 11. List of changes that took place in a single checklist.

Case Studies

In order to test the usability of this program in extracting names and data from a variety of sources, as well as to identify interesting analyses and visualizations that should be incorporated into the software itself, I developed three use cases. All case studies are based on publicly-available datasets.

Case Study 1: AmphibiaWeb

Introduction. AmphibiaWeb is an online taxonomic resource for global amphibian species that includes a comprehensive taxonomy. The current version of the taxonomy was first published in March 2012 (Blackburn, Cannatella, and Wake 2017). AmphibiaWeb has archived this taxonomy on a monthly basis since October 2012 (https://github.com/AmphibiaWeb/taxonomy-archive). These archives contain

comma-separated value (CSV) files that include a list of all recognized species and other information. Each recognized species also has an AmphibiaWeb ID that does not change when a species is renamed.

Principle questions. AmphibiaWeb's detailed checklist history allowed me to extract synonymy information easily and to quantify how this checklist has changed over four years. I focused on the following questions:

1. How stable have names been over this time period? How does that change once we take synonymy into account?
2. Which new species were added to the taxonomy over this time period? Why were species deleted?
3. How do the different synonymy inference tools available in SciNames compare with each other?

Methods. I started by importing the 52 checklists that AmphibiaWeb has made available between October 2012 and January 2017 and determined the stability of names that had remained unchanged between each checklist and the October 2012 checklist to determine the degree of similarity using a Jaccard index as described above.

I used the three available techniques in SciNames to infer synonymies to determine which technique provided the maximum number of synonyms. I then recalculated the similarity of names between each checklist and the October 2012 checklist, treating synonyms as identical. I also compared names between the first and last checklist, treating synonymous names as identical. I used publicly available data to determine when these names had been described in order to determine what proportion of them had been described since 2013, representing

new discoveries rather than previously described species that had not been recognized in October 2012.

Results. The AmphibiaWeb checklist recognized 7,039 species across 537 genera in October 2012. From there, it expanded to recognize 7,614 species (575 species, 8.2% increase) across 548 genera (11 genera, 2.0% increase) in January 2017. This change is extremely gradual – every checklist shared at least 96% of names with the checklist before it – suggesting that we see no sharp changes in the species that have been incorporated into this checklist (Figure 12). However, they add up: the last checklist in this project shared only 6635 species names (82.7% of combined names) with the first checklist. There are 979 binomial names recognized in the last checklist that are not present in the first checklist.

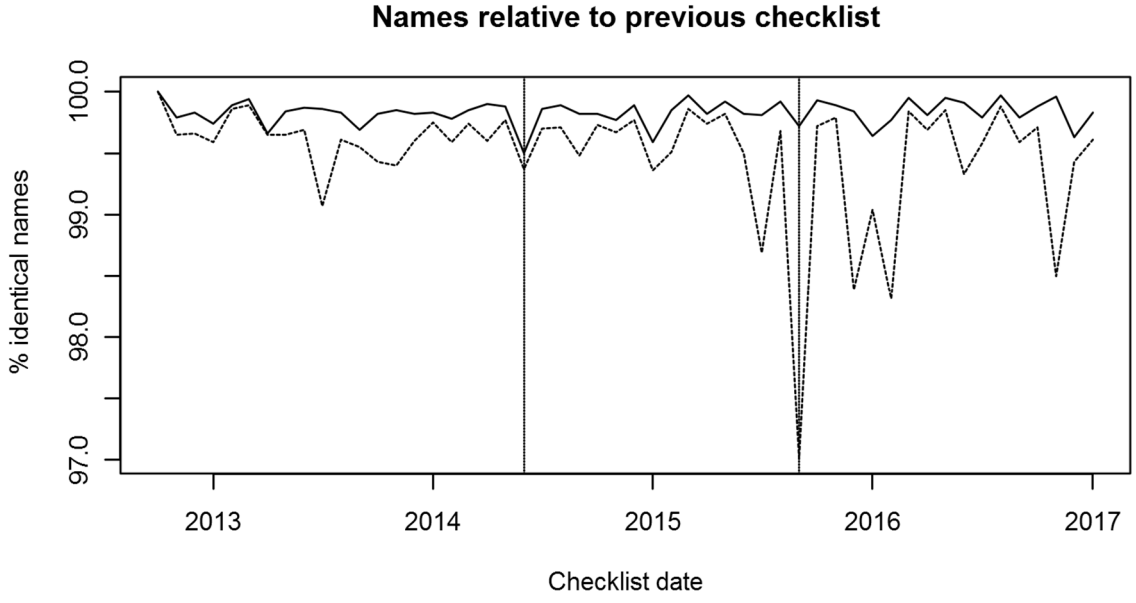


Figure 12. Names relative to the immediately previous checklist for AmphibiaWeb. The solid line indicates the similarity of name clusters while the dotted lines indicate the similarity of names between pairs of checklists. The two vertical lines mark dates when the composition of species in the checklist changed without affecting synonyms (June 2013) and when species in the checklist were synonymized, changing names without changing the composition of species in the checklist (September 2015).

Of these 979 missing binomial names, some are likely synonyms of previously recognized names. To determine how many there were, I used three different options for inferring renames from SciNames:

1. SciNames can identify when an inferred "addition" and "deletion" share an identifier in a specified column, and replace them with a "rename" from one name to the other. AmphibiaWeb provides just such a column called "aweb_uid", which is a species-specific identifier that does not change when a name is synonymized. This algorithm resulted in 496 synonyms being identified, of which five were duplicates, giving 491 unique synonyms.
2. A variation of this algorithm is to look at all binomial name additions and deletion that have taken place in this project, and determine if the name being added or deleted is associated with any other name through a shared identifier. Using the "aweb_uid" column, SciNames identified 987 synonyms, of which 57 synonyms were duplicates, resulting in 930 unique synonyms.
3. AmphibiaWeb also provides a "synonymies" column, in which comma-separated synonyms are included. These may be names not otherwise recognized in this dataset, providing us with a large store of names to include. Extracting these appeared to result in 3,670 synonyms, but most of these were duplicates, since the value of the synonymies column often did not change from checklist to checklist. There were only 140 unique synonyms generated by this method.

I used all three algorithms to identify synonymous species, and then plotted the similarity between each checklist and the October 2012 checklist over time. This plot shows how the similarity changes when taking synonymies into effect (solid line) as well as when ignoring them, and comparing name lists directly (dashed line). We see a small but significant difference that increases over time, to a

minimum of 82.8% similarity when ignoring synonymy information and 91.6% when including them (Figure 13).

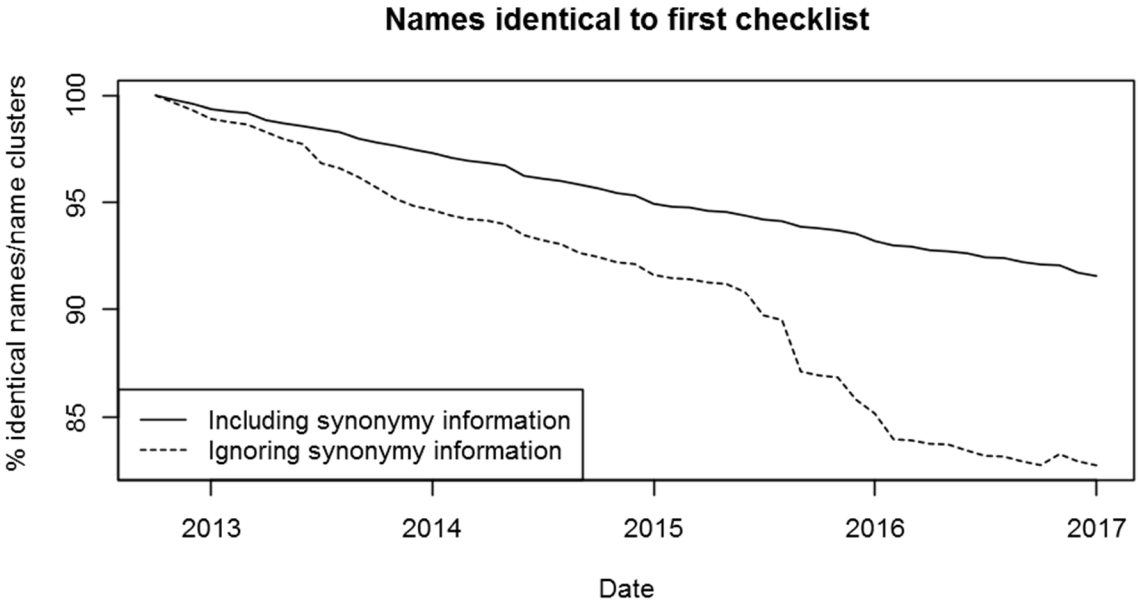


Figure 13. Names relative to the first checklist (October 2016) for AmphibiaWeb.

This graph is likely more useful than the similarity of names relative to the immediately previous checklist (Figure 12), both including synonymies (solid line) as well as when ignoring them (dashed line). While difficult to interpret, this graph can be useful to identify checklists where the contents of the checklist were modified (such as the first vertical line, June 2013) as well as cases where the names in the checklist were renamed with far fewer changes in the actual content of the checklist (such as the second vertical line, September 2015). This might be of value to managers of taxonomic checklists, and could be useful in identifying times of relative stability and instability.

Relative to the October 2012 checklist, the January 2017 checklist has 979 additional species names and 404 fewer species names. Many of these changes involved synonymous names; after eliminating those using the synonymies

identified above, I was left with 635 species names added (64.9% of name-only additions) and 4 deleted (1% of name-only deletions).

The four deleted species were:

1. *Crossodactylus bokermanni*: synonymized with *C. trachystomus* (Pimenta, Caramaschi, and Cruz 2015).
2. *Xenophrys kempii*: taxonomy appears to be in dispute; may refer to the same species as *Philautus kempii* (Frost 2017).
3. *Amietia dracomontana*: synonymized with *A. quecketti* (Channing and Baptista 2013).
4. *Polypedates iskanderi*: Appears to be a misspelling of *Polypedates iskandari*, which has been recognized through the entire period of this checklist.

Thus, all four deletions are junior synonyms of other names, and were excluded only because of missing synonym information.

With 635 species names added, there were too many to examine individually, but I used the TaxRefine online tool (previously written by me, see http://blog.vertnet.org/post/56169017224/taxonomic-validation-vaidya) with OpenRefine (Ham 2013) to quickly determine the description date for 320 names. While these ranged from 1824 to 2016, the vast majority of them – 88.1% – were described after 2010. This is also reflected in the summary statistics: the first quartile is 2013, the median is 2014, and the mean is 2005. This suggests that while some of these additions are previously described species that were subsequently unrecognized, the vast majority of them are recently discovered and described species that were previously unknown.

Case Study 2: Reptile Database

Introduction. The Reptile Database (Uetz 2016) is a global catalogue of living reptile species and their classification. It is far more typical of taxonomic checklists available online than AmphibiaWeb was: checklists are archived at irregular intervals, in different formats, with different levels of detail. This case study tests SciNames' ability to import different types of files to provide a single chronology of taxonomic changes.

Principle questions. Without synonymy information, SciNames can only provide basic information on how stable the list of recognized species has remained. Therefore, we can only ask the two basic questions SciNames can provide answers to:

1. How stable has the Reptile Database been over the last decade?
2. Which species have been added to and deleted from the checklist over this time period?

Methods. The Reptile Database input files were available in three file formats, all of which could be opened in Microsoft Excel:

- Four checklists between October 2006 and December 2014 available as compressed tab-delimited files.
- Seven checklists between March 31, 2013 and March 23, 2015 available as compressed Microsoft Excel XLS files.
- Five checklists between August 12, 2015 and December 24, 2016 available as Microsoft Excel XLSX files.

I used Microsoft Excel to load all these files and save them as UTF-8 encoded comma-separated values (CSV) files. Some of them were missing column headers,

which I inserted. I then imported these into SciNames. Two different name extractors were necessary: thirteen checklists stored binomial names in a single column, while three checklists stored the genus and specific epithets separately.

Once all the checklists had been loaded into SciNames and all the names extracted, I followed the same procedure as in the AmphibiaWeb case study to measure the similarity between each checklist and the first checklist (October 2006), and to identify the species added and deleted over the course of the entire checklist.

Results. The Reptile Database has a similar degree of name stability as AmphibiaWeb when comparing a similar time frame: the December 2016 checklist has 81.2% similarity with the March 2013 checklist. However, the similarity is much lower at 54.2% when compared with the October 2006 dataset, suggesting that extensive changes took place from 2006 to roughly 2014, with the rate of change diminishing slightly since then. As with AmphibiaWeb, both these numbers would likely be higher once we take synonymies into account; however, this dataset does not contain synonymy information for all checklists.

I identified 3,836 species that had been added to the Reptile Database since October 2006, and 1,805 species that had been deleted over this time period, but many of these are likely species being renamed to their synonyms.

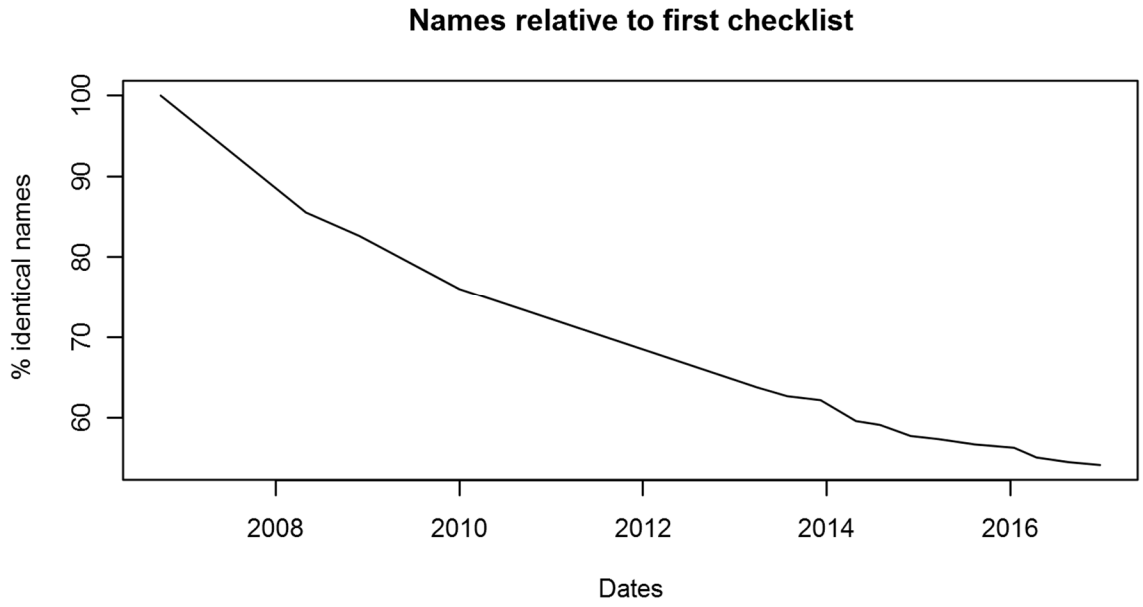


Figure 14. Similarity to the first checklist for the Reptile Database from 2006 to 2016.

Many files required extensive pre-processing before they could be loaded into SciNames. SciNames might benefit from better pre-processing tools that would allow messy data to be cleaned up and included into a project. However, other tools, such as Excel and R, are already widely used to clean messy data. One tool in particular, OpenRefine (Ham 2013), has been created specifically to deal with a variety of input formats, character encodings and other complex imports. Once these data have been cleaned and organized, they can be published separately from the checklist.

A key distinguishing feature of this dataset was the inclusion of Unicode control characters, specifically the start of heading (Unicode 1), vertical tab (Unicode 11), end of medium (Unicode 25), the group separator (Unicode 29) and the device control character (Unicode 144) in the included data. Note that some of these may have been introduced by the conversion in Excel, and might originally have been represented by other characters. Unfortunately, these characters cannot be

included in an XML file, so projects containing these characters cannot be loaded into Java. Filtering such characters automatically on import may result in data being lost – in the case of the Reptile Database, for instance, these characters are used to separate subspecies names, which may be valuable information. For this case study, I removed these by hand, but future versions of SciNames might need to incorporate automatic encoding and decoded of such entities.

Case Study 3: Data reconciliation with CITES

Introduction. The Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) is an international agreement that aims to regulate trade in certain species between 183 parties. According to its website, it currently covers roughly 5,600 species of animals and 30,000 species of plants. It divides species into three categories:

1. Appendix I: species threatened with extinction, trade permitted only in exceptional circumstances.
2. Appendix II: species not necessarily threatened with extinction, but in which trade must be controlled in order to avoid overutilization.
3. Appendix III: species protected by one or more CITES countries individually, but not by the convention in general.

Methods and Results. I attempted to reconcile names from AmphibiaWeb and the Reptile Database against the names in CITES in June 2017 by directly matching names across these two datasets, and by comparing their performance to the reconciliation tool in SciNames. Since the reconciliation tool was able to draw on information on synonyms (for AmphibiaWeb) and previously recognized names (in both checklists), I wanted to determine how much better SciNames would be at matching names when using name and synonym information from multiple

checklists, as compared to the far simpler strategy of simply matching identically spelled names. I used R to match identical names.

There were 164 amphibian species listed in CITES in June 2017. R was able to match 161 names (98.2%) directly against the January 2017 checklist, and 157 names (95.7%) against the October 2012 checklist. The degree of the match is therefore very high as long as an approximately contemporary checklist is chosen. SciNames outperformed both of these methods, matching 163 names (99.4%) when using names from the entire project, missing only *Andinobates virolensis*.

Without synonym information, the Reptile Database performed less well. CITES contained 876 reptile species. R was able to match 824 names (94.1%) against the December 2016 Reptile Database, while SciNames was able to match 851 names (97.1%). The difference was the result of SciNames being able to match names from different datasets: most names were first added in March 2013, but 33 names (3.9% of matched names) were first added in other checklists between December 2014 and April 2016.

Discussion and conclusions

Use in measuring taxonomic stability or as a reconciliation tool. The first two case studies included in this chapter as well as Chapter 2. The tempo and mode of the taxonomic correction process in North American Birds over the last 127 years and Chapter 3. The components of the taxonomic discovery process and the effect of abundance in the ongoing discovery of North American freshwater algae demonstrate the value of SciNames in measuring taxonomic stability. Where synonymy information is not present, it can provide a simple measure of name similarity over the checklist series (as in Case Study 2: Reptile Database). If synonym information is available, it can additionally provide a measure of name

cluster or nominal concept similarity that ignores changes in names where the two species names are synonymous with each other (as in Case Study 1: AmphibiaWeb). Synonymy information can be inferred automatically from a shared identifier or a column of synonyms, which many checklists already contain; they can also be added manually, such as by identifying additions and deletions involving species with similar specific epithets or importing a list of synonyms from another checklist or taxonomic database. Visualizing checklist similarity over time may be useful in identifying periods when the taxonomy of the group changed dramatically or to identify periods of relative stability, where analyses might be carried out with less work needed to reconcile different taxonomic circumscriptions.

While SciNames also appears to be useful as a name reconciliation tool, the included reconciliation case study (Case Study 3: Data reconciliation with CITES) shows that it might not provide a large improvement over simply matching scientific names, or by identifying a taxonomic checklist that can bridge the data being aggregated effectively. Given the challenges in reconciling names, scientists preparing data for analysis may opt to simply ignore data that cannot be reconciled instead of working to perfect the reconciliation, in which case SciNames might be unnecessary. Instead, SciNames could be used to prepare synonymy and circumscription change information from a series of checklist, which could then be used by software designed expressly to clean, reconcile and aggregate datasets, such as OpenRefine (Ham 2013), taxize (Chamberlain and Szöcs 2013) or GNparser (Mozzherin, Myltsev, and Patterson 2017).

One SciNames feature I have missed in other tools is the ability to synthesize a scientific name from multiple columns in a biodiversity data file, and to describe that synthesis in a standard format (see *Text processing of scientific names*). This feature might be integrated into other biodiversity data management and analysis

tools, which – if found broadly useful – could see this system being standardized and used to prepare datasets for conversion into standard biodiversity data storage formats, such as Darwin Core (Wieczorek et al. 2012).

Use as an editor of taxonomic changes. While all three test cases included in this chapter focused on checklist series, datasets and synonymy, both Chapter 2. The tempo and mode of the taxonomic correction process in North American Birds over the last 127 years and Chapter 3. The components of the taxonomic discovery process and the effect of abundance in the ongoing discovery of North American freshwater algae were based on datasets of taxonomic changes created in SciNames. For Chapter 2. The tempo and mode of the taxonomic correction process in North American Birds over the last 127 years, I created the taxonomic changes as entries in plain text files, which I later imported into SciNames for validation and testing. For Chapter 3. The components of the taxonomic discovery process and the effect of abundance in the ongoing discovery of North American freshwater algae, I annotated changes directly in SciNames. Having a tool for categorizing, validating and inferring taxonomic changes was extremely valuable, and this dissertation would not be possible without it. SciNames' flexible design allowed me to add new features as needed for particular analyses, such as the analysis of reverting lumps and splits in Chapter 2. The tempo and mode of the taxonomic correction process in North American Birds over the last 127 years or the analysis of species being renamed among currently recognized genera in Chapter 3. The components of the taxonomic discovery process and the effect of abundance in the ongoing discovery of North American freshwater algae. Since SciNames has been released under an open source license, other scientists will be able extend it as needed for their own analyses of checklist series data.

One feature that I found particularly valuable was the distinction between explicit and implicit changes. As implicit changes are annotated into renames, lumps, or splits, they are automatically removed as implicit changes. Erroneous explicit changes result in an implicit change being created to undo them, allowing them to be detected immediately. The validation tool is also valuable in detecting inconsistencies such as duplicate additions or deletions or a name being used multiple times in the same change.

Use in managing taxonomic checklists. While much of my dissertation focuses on taxonomic changes, SciNames could also be used to manage taxonomic checklists. Published checklists should include information on how each new version of the checklist differs from the previous version, including a list of changes, justifications for each change and citations to the relevant literature. By using SciNames' XML format, checklists could provide change information in a machine-readable format while also providing justifications and citations for the change in a human-readable format. Creating these datasets in SciNames would ensure that every change was accounted for. SciNames could be extended so that new changes were automatically checked against previous ones, warning checklist managers when they were reverting a change that was made previously and reminding them of the justification and citations for the previously made change. This would allow databases of taxonomic changes and circumscriptional changes to be produced as a byproduct of the curation process necessary for managing any checklist.

Biodiversity data integration projects that use checklists to reconcile data sometimes need to modify those checklists, whether because of a discrepancy between the checklist and the data they are integrating, because of changes made in the relevant taxonomy since the checklist, or because of different taxonomic opinions between the integrators and the authors of the checklist. SciNames would

allow these project-specific changes to be recorded and annotated, along with justifications and citations, and will hopefully encourage these project-specific checklists to be published themselves, providing more datasets of taxonomic opinions for researchers to study.

I focused on checklist series in my dissertation, as these allowed me to quickly identify changes within a stable taxonomic view over a long period of time. However, SciNames could also be used to compare multiple taxonomic views across a single period of time, such as between contemporary taxonomists who use different philosophies or between different organizations that create checklists. This more closely resembles the two major databases of taxonomic circumscriptions: Avibase, which reconciles different versions of bird taxonomic checklists from around the world (Lepage 2017; Lepage, Vaidya, and Guralnick 2014), and the *Flora of the Southern and Mid-Atlantic States* (Weakley 2015), which is a conventional taxonomic checklist that additionally provides information on how circumscriptions within the checklist differ from those in previous works covering the same geographical area.

Future development. While the data provided by SciNames can be easily visualized in statistical software such as R, a valuable goal for the next version of SciNames would be for the visualization to occur within the application itself. This would allow for interactive graphs that could be zoomed into, with points on the graph to jump directly to the appropriate dataset.

To improve the ability of SciNames to extract valuable information from checklist series, the most immediately useful addition would be better inferencing tools. Potential renames could be identified automatically by identifying species being added to a checklist with an identical or similar specific epithet to a species being deleted from the same checklist. Names could be validated against Taxonomic

Name Resolution Services such as the one provided by iPlant (Boyle et al. 2013). Synonyms could be queried directly from online nomenclatural databases. Changes in which a species had been renamed to a subspecies or where a subspecies had been renamed to a species could be automatically identified as a lump or a split respectively. Ultimately, SciNames might be able to import a dataset, immediately generate a large number of inferred changes along with links to the evidence used to make those inferences, and then provide an interface for a human operator to rapidly verify those inferences.

SciNames could also be extended to import and export data in different formats to allow for intercommunication between different approaches to managing and sharing taxonomic changes and concepts. In recent years, Franz and colleagues have been working on a tool for reasoning over known relationships between taxonomic concepts (Nico M Franz et al. 2016; N. Franz and Cardona-Duque 2013; M. Chen et al. 2014), while other scientists have been working on an OWL ontology representation of taxonomic change (Tuominen and Laurenne 2011; Laurenne et al. 2014). A tool like SciNames might be well-placed to allow data used by one group to be represented as a set of taxonomic changes, stored in its native XML representation, and converted into the other representation for use in other software. As a flexible and sophisticated tool for managing checklist series data and annotating resulting taxonomic changes, it is well-suited to serve as an intercommunication platform for tying together different ideas of how taxonomy can be modeled, how these models could be used to improve biodiversity data integration, and what biological questions we could answer with the next generation of taxonomic change datasets.

Supplementary Materials

The datasets used in these case studies are based on published taxonomic checklists. They have been uploaded to Figshare at https://figshare.com/s/012ed914fdf4a95d4a5a to help with reproducibility. Once this chapter has been published, and in consultation with the authors of the taxonomic checklists they are based on, these data will be made publicly available with their own DOI.

Appendix 1: Inferring changes

SciNames provides several options for inferring changes. Some of these require a dataset column to be specified. Note that all inferred changes are guesses, and should be verified before being included in the project. The four currently implemented inference methods are:

- *Find renames using identifiers in additions and deletions* (requires column containing unique identifier): checklists sometime use unique identifiers to track names regardless of synonymy, so that a name can be moved from one genus to another but retain the same identifier. This inference method looks for a name being deleted from a checklist that shares the same value in the provided identifier column with another name being added; if found, it infers that the first name was renamed to the second name in the dataset in which the addition and deletion took place.
- *Find renames using identifiers in data* (requires column containing unique identifier): checklists sometime use unique identifiers to track names regardless of synonymy, so that a name can be moved from one genus to another but retain

the same identifier. This inference method looks for all names in the entire dataset that share the same identifier value in this column; all such names are inferred to be related through "rename" changes. It attempts to identify which dataset the "rename" change should be placed in, but may be unsuccessful; inferred changes should therefore be reviewed and added manually if a truly unknown synonymy is found.

- *Find renames using a synonym column* (requires column): checklists often contain a column with a list of synonyms for the specified species. These are usually separated by commas (,), semicolons (;) or pipes (|). This method infers "rename" changes that correspond to each of these synonyms.
- *Genus reorganization from renames*: This generates lists of synonyms where species have moved between two genera that are both recognized in the specified dataset. The inferred results duplicate existing synonymies, and so should not be used – it was specifically developed to carry out a single analysis for Chapter 3. The components of the taxonomic discovery process and the effect of abundance in the ongoing discovery of North American freshwater algae.

Appendix 2: Change filters

The need for change filters was motivated by a need to eliminate post-1980 changes in order to analyze AOU Checklists (see Chapter 2. The tempo and mode of the taxonomic correction process in North American Birds over the last 127 years). Change filters are a very low-level interface, allowing changes to be ignored from all analyses. The following change filters are currently included in SciNames:

- "ignoreError": Filters out all changes having the type "error".
- "ignoreIgnored": Filters out all changes that have a property of "ignored" with a value of "yes".

- "skipChangesUnlessAddedBefore" (parameter: "year"): skips all changes unless they include a name cluster that was added before the specified year.

Appendix 3: Validation tools

A large project may contain multiple datasets, checklists and changes, and a single incorrect change or misinterpreted dataset may lead to inaccurate results. To facilitate validation, SciNames contains a tool that runs prewritten tests to identify suspicious or inconsistent data at different levels:

1. At the dataset level, SciNames tests:
    a. Whether all rows could be mapped to a scientific name,
    b. Whether any of its changes contradict each other, such as a name added in one change that is deleted in another change in the same dataset, and
    c. Whether any of its changes have no effect over the previous dataset, such as by adding a taxon that is already recognized or deleting a name that was not previously recognized.
2. At the change level, SciNames:
    a. Checks that all changes are of one of the five recognized change types: "addition", "deletion", "lump", "split", "rename" or "error",
    b. Checks whether changes of type "addition" do in fact add taxa and changes of the type "deletion" do in fact delete taxa,
    c. Validates changes of type "lump" and "split" by checking that multiple taxa are lumped into a single taxon and that a single taxon is split into multiple taxa,

d. Checks whether a single change involves multiple synonyms of the same name, and

e. Checks whether names can be expressed in ASCII. If a project includes non-English languages as names, this is likely to provide spurious results, and so the priority for this notice has been set low.

3. At the name clusters level, SciNames carries out basic consistency checking, making sure that the same name has not been added to multiple name clusters.

CHAPTER 5. DISCUSSION AND CONCLUSIONS

The studies included in this dissertation show that questions about the taxonomic correction process can be tackled on very different time scales, about different types of relationships, in different stages of taxonomic completion and on very different types of living organisms with the help of taxonomic checklists. I found a strong temporal pattern in species redescription in a checklist of North American birds, with extensive lumping peaking in the 1930s and 1940s replaced almost entirely with extensive and accelerating splitting after the 1980s. There was no clear temporal pattern in a checklist of North American freshwater algae, but lumping and splitting have occurred throughout this checklist and show no sign of decreasing. Splitting in freshwater algae does not appear to be accelerating, suggesting that accelerating splitting may be unique to North American vertebrate taxa. I found that nearly 20% of biodiversity records in one freshwater algae dataset were associated with species known to have multiple circumscriptions, showing that taxonomic redescription could affect interpretation of a significant number of biodiversity records.

Over the longer time period covered by the bird checklist, I found that 74% of species recognized today had never been lumped or split. Over the shorter time period of the freshwater algae checklist, I found that larger genera had larger numbers of lumps and splits, that species being split and lumped are more abundant than other species, and that when moving species from one currently-recognized genus to another, both the source and destination genus tend to be more

speciose than other genera. This suggests that smaller genera and species known from fewer observations might be undergoing less redescription than larger genera and species known from many observations, possibly because they are closer to complete and so require fewer corrections. This may also reflect a bias towards larger genera and better sampled species in the taxonomic redescription process.

By focusing my study on taxonomic changes within taxonomic checklists, I sidestepped the broader question of how often taxonomic changes are proposed, what proportion of proposed changes are ignored or disputed, and how long it takes these proposed changes to become accepted. As a result, my findings underestimate both the amount of redescription proposed for these groups and the total taxonomic effort being put into these groups, possibly by a large amount. To identify this, studies carried out using checklist-based approaches could be contrasted with those that examine taxonomic proposals directly (such as Sangster 2009, 2014). On the other hand, my findings apply directly to the primary way in which scientists use taxonomic names, and so measure a more practically useful effect of taxonomic change than other approaches do.

The checklist-based approach has many benefits in terms of reducing the amount of work and taxonomic knowledge necessary to quantify taxonomic change. In particular, each successive checklist shares many of its names and circumscription with the previous checklist, allowing a scientist annotating those changes to focus only on names that have changed. However, this focus means that circumscription changes not visible as name changes may be ignored by the

checklist, such as the movement of subspecies from one species to another (which would change both circumscriptions) or a change in geographical distributions in which species found in one location are reclassified into another species. This model of checklist-based analyses should therefore be compared with approaches that rely on taxonomic experts providing articulations between taxonomic circumscriptions (Nico M Franz et al. 2016), or compared to taxonomic circumscriptions expressed in geographical space (by drawing range maps for different species) or in morphospace, such as by using the output of a tool such as ETC (Cui et al. 2016) or described using an anatomy ontology (Mullins et al. 2012).

While the model I used was sufficient to describe the taxonomic changes I observed in these datasets, it may be possible to simplify it further. Given that the product of descriptions and redescriptions are both novel taxonomic circumscriptions, it may be that the only distinction between these processes is whether the novel circumscription contains an existing type specimen (making it the redescription of an existing, named species) or not (making it an original description). In this model, the key distinction between North American birds and freshwater algae is that the former has so many species and subspecies-level descriptions from the 18th and 19th centuries that it is unlikely that a novel circumscription will contain no type specimens, and so newly generated circumscriptions are largely redescriptions of existing (usually subspecific) taxa. In freshwater algae, by contrast, most novel circumscriptions contain no previously named type specimens, and so need to be described as new taxa. This simpler model

might be more flexible than the one used in my dissertation, but might require specific knowledge about which type specimen each type is tied to.

Apart from such vertical analyses of checklists over time, checklists may also be analyzed horizontally, comparing checklists produced by different authors at the same time. The differences between these checklists would not represent a difference in evidence, assuming all authors are equally well-informed, but differences in interpreting that evidence to determine taxonomic boundaries. For example, Amphibian Species of the World (Frost 2017) recognizes *Ambystoma subsalsum* as a distinct species, while AmphibiaWeb (Blackburn, Cannatella, and Wake 2017) does not, possibly considering it to be a synonym of *Ambystoma velasci*. While vertical series are useful in measuring how quickly taxonomy changes, horizontal series could be a useful source of data in understanding taxonomic disagreements: how often they occur, how long they persist, and whether better technologies and wider dissemination of scientific outputs through the internet make such disagreements rarer or more quickly resolved.

One unanswered question at the end of this dissertation is how universal the observed patterns really are. Do they extend beyond North American taxa? Are they specific to freshwater algae and birds in some way? If I could start my dissertation over, I would have started by building the software tool first, and then using it to build more datasets that could answer a smaller number of questions over a larger number of geographically and taxonomically disparate checklists. Encouraging other scientists to undertake that research would involve establishing important

questions capable of being answered by a feasibly small number of checklists, and ensuring that the software tools available to build, analyze, validate, and publish are available and easy-to-use. Some possible questions include:

(1) Is splitting increasing across all vertebrate groups, as seen in primates and amphibians (Isaac, Mallet, and Mace 2004; Padial and de la Riva 2006)? Is there evidence for accelerated splitting as we see in North American birds?

(2) Is there a clear difference in species lump and split rates between smaller and larger genera? Is there any other evidence of a higher taxonomic bias in species lump and split rates?

(3) How many species have moved between genera over time? How often are genera split from or lumped into other genera, and how often are those changes reversed? Can we use similarity measures to compare the circumscription of genera over time, or do we need a more sophisticated measure that takes into account species-level lumping and splitting in the changing circumscriptions of genus-level entities?

Checklists today are generally treated as a single, definitive list: when errors are found or the taxonomy is updated, the checklist is replaced by a new list. Some checklists make previous versions of their taxonomy available for users to download. Even when checklists keep detailed records on which taxa changed and how, that data is not always easy for users to download and use. With my software tool, checklists could be annotated by checklist managers with detailed information on why each change was made, including citations to literature where relevant.

These could be used to produce detailed changelogs that are distributed with the checklist, or provided to the user in an XML file that contains both multiple versions of a checklist as well as annotated lists of changes between them.

The software tool I built, SciNames, could also potentially change how taxonomic checklists are used by end-users. When users use a checklist for reconciliation or aggregation, they often need to customize that checklist for their project: to fix minor corrections, to allow for alternate taxonomic interpretations, or to incorporate taxonomic changes made since the original publication of the checklist. The original checklist could be loaded in SciNames and project-specific changes made to it annotated carefully, allowing those changes to be stored with the integrated data, improving reproducibility. This would also automatically reconcile concepts between the project-specific checklist and the original. In some cases, a data integration project will use multiple checklists from different sources to handle different taxonomic groups. SciNames cannot currently support this type of integration, but this feature could be added if it would be useful.

My dissertation shows that the taxonomic correction process is proceeding strongly in several different taxonomic groups, and has been for decades. While this is unsurprising, what may be surprising is the enormous impact on biodiversity data and its meaning. In one dataset of North American freshwater algae, almost 20% of records are ambiguous in terms of circumscription within just one checklist. As taxonomy proceeds, it affects how biodiversity data should be interpreted and how easily checklists can be reconciled. It also reflects the health of an important

aspect of taxonomy, that of testing and correcting previously published circumscriptions. My work provides a simple and straightforward method for extracting much of the core of that process – synonyms, lumps, and splits – from series of taxonomic checklists, and a software tool for facilitating this process.

Taxonomic checklists are ubiquitous in biodiversity informatics: any large-scale project integrating data from different data resources is based on lists of curated names, whether it is a custom checklist developed and maintained by a single project like the AmphibiaWeb Taxonomy (Blackburn, Cannatella, and Wake 2017) or the GBIF Backbone Taxonomy, a massive synthesized checklist containing over 3 million accepted names and 2.2 million synonyms from 54 taxonomic sources used as a taxonomic backbone for the 855 occurrence records in GBIF (GBIF Secretariat 2017). My work shows how these checklists can be used to understand taxonomic practice, to document the relationships between taxa in these different checklists, to determine why taxonomic checklists differ from each other over time and to work towards synthesizing all biodiversity knowledge under a single, universal, thoroughly validated index.

REFERENCES

Agapow, Paul-Michael, Olaf R P Bininda-Emonds, Keith A Crandall, John L Gittleman, Georgina M Mace, Jonathan C Marshall, and Andy Purvis. 2004. "The Impact of Species Concept on Biodiversity Studies." *The Quarterly Review of Biology* 79 (2): 161–79. http://www.jstor.org/stable/10.1086/383542.

Aldrich, John W. 1946. "Speciation in the White-Cheeked Geese." *Wilson Bulletin* 58 (2): 94–103. http://sora.unm.edu/node/126683.

Alroy, John. 2002. "How Many Named Species Are Valid?" *Proceedings of the National Academy of Sciences* 99 (6). National Academy of Sciences: 3706–11. doi:10.1073/pnas.062691099.

American Ornithologists' Union. 1886. *The Code of Nomenclature and Check-List of North American Birds Adopted by the American Ornithologists' Union; Being the Report of the Committee of the Union on Classification and Nomenclature.* Cambridge: John Wilson and Son. http://dx.doi.org/10.5962/bhl.title.1538.

———. 1931. *Check-List of North American Birds, 4th Edition*. Lancaster, Pa.: The Union,. doi:10.5962/bhl.title.6394.

———. 1983. *Check-List of North American Birds*. 6th ed. Lawrence, Kansas. doi:10.5962/bhl.title.50892.

———. 1998. *Check-List of North American Birds: The Species of Birds of North America from the Arctic through Panama, Including the West Indies and Hawaiian Islands*. 7th ed. Washington, D.C.: American Ornithologists' Union. http://www.worldcat.org/isbn/189127600X.

American Ornithologists' Union. Committee on Classification and Nomenclature. 1998. *Check-List of North American Birds: The Species of Birds of North America from the Arctic through Panama, Including the West Indies and Hawaiian Islands*. 7th ed. Washington D.C: The Union. https://www.worldcat.org/title/check-list-of-north-american-birds-the-species-of-birds-of-north-america-from-the-arctic-through-panama-including-the-west-indies-and-hawaiian-islands/oclc/610812528&referer=brief_results.

Audubon, John J. 1835. "Hutchins's Goose." In *Ornithological Biography, or an Account of the Habits of the Birds of the United States of America ; Accompanied by Descriptions of the Objects Represented in the Work Entitled The Birds of America, and Interspersed with Delineations of American Scenery a*, 3:526–28. Edinburgh: Adam & Charles Black. doi:10.5962/bhl.title.48976.

Banks, Richard C, Carla Cicero, Jon L Dunn, Andrew W Kratter, Pamela C Rasmussen, J. V. Remsen, James D Rising, and Douglas F Stotz. 2004. "Forty-

Fifth Supplement to the American Ornithologists' Union Check-List of North American Birds." *The Auk* 121 (3): 985–95. doi:10.1642/0004-8038(2004)121[0985:FSTTAO]2.0.CO;2.

Barrowclough, George F., Joel Cracraft, John Klicka, and Robert M. Zink. 2016. "How Many Kinds of Birds Are There and Why Does It Matter?" *PLOS ONE* 11 (November): e0166307. doi:10.1371/journal.pone.0166307.

Barve, Vijay. 2014. "Discovering and Developing Primary Biodiversity Data from Social Networking Sites: A Novel Approach." *Ecological Informatics* 24 (November): 194–99. doi:10.1016/j.ecoinf.2014.08.008.

Bebber, Daniel P, Francis H C Marriott, Kevin J Gaston, Stephen A Harris, and Robert W Scotland. 2007. "Predicting Unknown Species Numbers Using Discovery Curves." *Proceedings of the Royal Society B: Biological Sciences* 274 (1618). The Royal Society: 1651–58. doi:10.1098/rspb.2007.0464.

Belore, Megan L, Jennifer G Winter, and Hamish C Duthie. 2002. "Use of Diatoms and Macroinvertebrates as Bioindicators of Water Quality in Southern Ontario Rivers." *Canadian Water Resources Journal* 27 (4): 457–84. http://www.tandfonline.com/doi/abs/10.4296/cwrj2704457.

Berendsohn, Walter G. 1995. "The Concept of 'Potential Taxa' in Databases." *Taxon* 44 (2): 207+. doi:10.2307/1222443.

Besse-Lototskaya, Anna, Piet F M Verdonschot, Michel Coste, and Bart Van de Vijver. 2011. "Evaluation of European Diatom Trophic Indices." *Ecological Indicators* 11 (2). Elsevier Ltd: 456–67. http://linkinghub.elsevier.com/retrieve/pii/S1470160X10001214.

Blackburn, David C., David C. Cannatella, and David B. Wake. 2017. "AmphibiaWeb Taxonomy." http://www.amphibiaweb.org/taxonomy/index.html.

Boyle, Brad, Nicole Hopkins, Zhenyuan Lu, Juan Antonio Raygoza Garay, Dmitry Mozzherin, Tony Rees, Naim Matasci, et al. 2013. "The Taxonomic Name Resolution Service: An Online Tool for Automated Standardization of Plant Names." *BMC Bioinformatics* 14 (1): 16+. doi:10.1186/1471-2105-14-16.

Bray, Tim, and Google Inc. 2014. "The JavaScript Object Notation (JSON) Data Interchange Format." RFC 7159. Internet Engineering Task Force (IETF). https://tools.ietf.org/html/rfc7159.

California Academy of Sciences. 2017. "iNaturalist." California Academy of Sciences. http://www.inaturalist.org/.

"Catalogue of Life." 2017. Accessed May 8. http://www.catalogueoflife.org/.

Catesby, Mark. 1731. *The Natural History of Carolina, Florida and the Bahama Islands : Containing the Figures of Birds, Beasts, Fishes, Serpents, Insects, and Plants : Particularly, the Forest-Trees, Shrubs, and Other Plants, Not Hitherto Described, or Very Incorrectly Figured by Authors : Together with Their Descriptions in English and French : To Which, Are Added Observations on the Air, Soil, and Waters : With Remarks upon Agriculture, Grain, Pulse, Roots, &c. : To the Whole, Is Prefixed a New and Correct Map of the Countries Treated of*. Printed at the expence of the author, and sold by W. Innys and R. Manby, at the West End of St. Paul's, by Mr. Hauksbee, at the Royal Society House, and by the author, at Mr. Bacon's in Hoxton. http://www.worldcat.org/oclc/6327279.

Chamberlain, Scott A, and Eduard Szöcs. 2013. "Taxize: Taxonomic Search and Retrieval in R." *F1000Research*, September. doi:10.12688/f1000research.2-191.v1.

Channing, Alan, and Ninda Baptista. 2013. "Amietia Angolensis and A. Fuscigula (Anura: Pyxicephalidae) in Southern Africa: A Cold Case Reheated." *Zootaxa* 3640 (4): 501. doi:10.11646/zootaxa.3640.4.1.

Chen, Mingmin, Shizhuo Yu, Nico Franz, Shawn Bowers, and Bertram Ludäscher. 2014. "Euler/X: A Toolkit for Logic-Based Taxonomy Integration." *arXiv*, 1–8. http://arxiv.org/abs/1402.1992.

Chen, Xiang, Weiqi Zhou, Steward T A Pickett, Weifeng Li, Lijian Han, and Yufen Ren. 2016. "Diatoms Are Better Indicators of Urban Stream Conditions: A Case Study in Beijing, China." *Ecological Indicators* 60 (January). Elsevier Ltd: 265–74. http://linkinghub.elsevier.com/retrieve/pii/S1470160X15003726.

Chesser, R. Terry, Richard C. Banks, Kevin J. Burns, Carla Cicero, Jon L. Dunn, Andrew W. Kratter, Irby J. Lovette, et al. 2015. "Fifty-Sixth Supplement to the American Ornithologists' Union: Check-List of North American Birds." *The Auk* 132 (3): 748–64. doi:10.1642/AUK-15-73.1.

Chesser, R. Terry, Kevin J. Burns, Carla Cicero, Jon L. Dunn, Andrew W. Kratter, Irby J. Lovette, Pamela C. Rasmussen, et al. 2017. "Fifty-Eighth Supplement to the American Ornithological Society's *Check-List of North American Birds*." *The Auk* 134 (3). The American Ornithologists' Union : 751–73. doi:10.1642/AUK-17-72.1.

Chesser, R Terry, Richard C Banks, F Keith Barker, Carla Cicero, Jon L Dunn, Andrew W Kratter, Irby J Lovette, et al. 2011. "Fifty-Second Supplement to the American Ornithologists' Union Check-List of North American Birds." *The Auk* 128 (3): 600–613. doi:10.1525/auk.2011.128.3.600.

———. 2013. "Fifty-Fourth Supplement to the American Ornithologists' Union Check-List of North American Birds." *The Auk* 130 (3): 558–72.

doi:10.1525/auk.2013.130.3.1.

Chesser, R Terry, Richard C Banks, Carla Cicero, Jon L Dunn, Andrew W Kratter, Irby J Lovette, Adolfo G. Navarro-Sig?enza, et al. 2014. "Fifty-Fifth Supplement to the American Ornithologists' Union Check-List of North American Birds." *The Auk* 131 (4): CSi-CSxv. doi:10.1642/AUK-14-124.1.

Chesser, R Terry, Kevin J Burns, Carla Cicero, Jon L Dunn, Andrew W Kratter, Irby J Lovette, Pamela C Rasmussen, et al. 2016. "Fifty-Seventh Supplement to the American Ornithologists' Union Check-List of North American Birds." *The Auk* 133 (3). The American Ornithologists' Union: 544–60. doi:10.1642/AUK-16-77.1.

Collen, Ben, Andy Purvis, and John L. Gittleman. 2004. "Biological Correlates of Description Date in Carnivores and Primates." *Global Ecology and Biogeography* 13 (5): 459–67. doi:10.1111/j.1466-822X.2004.00121.x.

Constable, Heather, Robert Guralnick, John Wieczorek, Carol Spencer, and A. Townsend Peterson. 2010. "VertNet: A New Model for Biodiversity Data Sharing." *PLoS Biology* 8 (2): e1000309. doi:10.1371/journal.pbio.1000309.

Costello, Mark J., Simon Wilson, and Brett Houlding. 2012. "Predicting Total Global Species Richness Using Rates of Species Description and Estimates of Taxonomic Effort." *Systematic Biology* 61 (5): 871–83. doi:10.1093/sysbio/syr080.

Costello, Mark J, Marguerita Lane, Simon Wilson, and Brett Houlding. 2015. "Factors Influencing When Species Are First Named and Estimating Global Species Richness." *Global Ecology and Conservation* 4 (July): 243–54. http://www.sciencedirect.com/science/article/pii/S2351989415000748.

Costello, Mark J, Simon Wilson, and Brett Houlding. 2013. "More Taxonomists Describing Significantly Fewer Species per Unit Effort May Indicate That Most Species Have Been Discovered." *Systematic Biology*, April. Oxford University Press. doi:10.1093/sysbio/syt024.

Coues, Elliott. 1873. *A Check List of North American Birds. By Elliott Coues. A Check List of North American Birds. By Elliott Coues.* Salem [Mass.]: Naturalists' Agency. doi:10.5962/bhl.title.14114.

Cracraft, Joel. 1983. "Species Concepts and Speciation Analysis." In *Current Ornithology*, 1:159–187. Boston, MA: Springer US. doi:10.1007/978-1-4615-6781-3_6.

Cui, Hong, Dongfang Xu, Steven S Chong, Martin Ramirez, Thomas Rodenhausen, James A Macklin, Bertram Ludäscher, Robert A Morris, Eduardo M Soto, and

Nicolás Mongiardino Koch. 2016. “Introducing Explorer of Taxon Concepts with a Case Study on Spider Measurement Matrix Building.” *BMC Bioinformatics* 17 (1): 471. doi:10.1186/s12859-016-1352-7.

Darwin, Charles. 1859. *On the Origin of Species by Means of Natural Selection, Or, The Preservation of Favoured Races in the Struggle for Life /*. London : John Murray,. doi:10.5962/bhl.title.68064.

Eisenmann, Eugene, Dean Amadon, Richard C Banks, Emmet R Blake, Thomas R Howell, Ned K Johnson, Jr George H Lowery, Kenneth C Parkes, and Robert W Storer. 1973. “Thirty-Second Supplement to the American Ornithologists’ Union Check-List of North American Birds.” *The Auk* 90 (April): 411–19. https://sora.unm.edu/node/22371.

Eisenmann, Eugene, Burt L Monroe, Kenneth C Parkes, Lester L Short, Richard C Banks, Thomas R Howell, Ned K Johnson, and Robert W Storer. 1982. “Thirty-Fourth Supplement to the American Ornithologists’ Union Check-List of North American Birds.” *The Auk* 99 (3): 1CC–16CC. doi:10.2307/4085886.

Eisenmann, Eugene, Kenneth C Parkes, Richard C Banks, George H Lowery, Thomas R Howell, Burt L Monroe, Ned K Johnson, Robert W Storer, and Lester L Short. 1976. “Thirty-Third Supplement to the American Ornithologists’ Union Check-List of North American Birds.” *The Auk* 93 (4): 875–79. https://sora.unm.edu/node/22909.

Faria, Vicente V., Matthew T. McDavitt, Patricia Charvet, Tonya R. Wiley, Colin A. Simpfendorfer, and Gavin J. P. Naylor. 2013. “Species Delineation and Global Population Structure of Critically Endangered Sawfishes (Pristidae).” *Zoological Journal of the Linnean Society* 167 (1): 136–64. doi:10.1111/j.1096-3642.2012.00872.x.

Fourtanier, Elisabeth, and J Patrick Kociolek. 1999. “Catalogue of the Diatom Genera.” *Diatom Research* 14 (1). Taylor & Francis: 1–190. doi:10.1080/0269249X.1999.9705462.

Fourtanier, Elisabeth, and John Patrick Kociolek. 2003. “Addendum To ‘Catalogue of the Diatom Genera.’” *Diatom Research* 18 (2). Taylor & Francis Group: 245–58. doi:10.1080/0269249X.2003.9705590.

Franz, N.M., and R.K. Peet. 2009. “Perspectives: Towards a Language for Mapping Relationships among Taxonomic Concepts.” *Systematics and Biodiversity* 7 (1): 5–20. doi:10.1017/S147720000800282X.

Franz, Nico, and Juliana Cardona-Duque. 2013. “Description of Two New Species and Phylogenetic Reassessment of Perelleschus O’Brien & Wibmer, 1986 (Coleoptera: Curculionidae), with a Complete Taxonomic Concept History of

Perelleschus Sec. Franz & Cardona-Duque, 2013." *Systematics and Biodiversity* 11 (2). Taylor & Francis: 209–36. doi:doi: 10.1080/14772000.2013.806371.

Franz, Nico M, Naomi M Pier, DeeAnn M Reeder, Mingmin Chen, Shizhuo Yu, Parisa Kianmajd, Shawn Bowers, and Bertram Ludäscher. 2016. "Two Influential Primate Classifications Logically Aligned." *Systematic Biology* 65 (4). School of Life Sciences, PO Box 874501, Arizona State University, Tempe, AZ 85287, USA; nico.franz@asu.edu.: Oxford University Press: 561–82. http://sysbio.oxfordjournals.org/content/65/4/561.full.

Franz, Nico, Robert Peet, and Alan Weakley. 2008. "On The Use Of Taxonomic Concepts In Support Of Biodiversity Research And Taxonomy." In , 63–86. doi:10.1201/9781420008562.ch5.

Frodin, David G. 2004. "History and Concepts of Big Plant Genera." *Taxon* 53 (3). International Association for Plant Taxonomy: 753–76. doi:10.2307/4135449.

Frost, Darrel R. 2017. "Amphibian Species of the World: An Online Reference. Version 6.0." New York, USA: American Museum of Natural History. http://research.amnh.org/herpetology/amphibia/index.html.

Gaston, Kevin J., and Laurence A. Mound. 1993. "Taxonomy, Hypothesis Testing and the Biodiversity Crisis." *Proceedings of the Royal Society of London B: Biological Sciences* 251 (1331). http://rspb.royalsocietypublishing.org/content/251/1331/139.

GBIF Secretariat. 2017. "GBIF Backbone Taxonomy." doi:10.15468/39omei.

Gill, Frank B. 2014. "Species Taxonomy of Birds: Which Null Hypothesis?" *The Auk* 131 (2). The American Ornithologists' Union: 150–61. doi:10.1642/AUK-13-206.1.

Grant, T R. 1989. "Ornithorhynchidae." In *Fauna of Australia*. http://www.environment.gov.au/science/abrs/publications/fauna-of-australia/fauna-1b.

Grinnell, J. 1935. "Publication Reviewed: Catalogue of Birds of the Americas, Part VII by Charles E. Hellmayr." *The Condor* 37 (2): 90–92.

Groves, Colin P. 2014. "Primate Taxonomy: Inflation or Real?" *Annual Review of Anthropology* 43 (1): 27–36. doi:10.1146/annurev-anthro-102313-030232.

Guiry, M.D., and G.M. Guiry. 2017. "AlgaeBase." *World-Wide Electronic Publication, National University of Ireland, Galway*. http://www.algaebase.org.

Haffer, Jürgen. 1992. "The History of Species Concepts and Species Limits in

Ornithology." *Bulletin of the British Ornithologists' Club* 112A: 107–58. http://www.biodiversitylibrary.org/part/149170.

Ham, Kelli. 2013. "OpenRefine (Version 2.5). Http://openrefine.org. Free, Open-Source Tool for Cleaning and Transforming Data." *Journal of the Medical Library Association : JMLA* 101 (3). Medical Library Association: 233–34. doi:10.3163/1536-5050.101.3.020.

Hardisty, Alex, and Dave Roberts. 2013. "A Decadal View of Biodiversity Informatics: Challenges and Priorities." *BMC Ecology* 13 (1): 1–23. doi:10.1186/1472-6785-13-16.

Heller, Rasmus, Peter Frandsen, Eline D Lorenzen, and Hans R Siegismund. 2013. "Are There Really Twice as Many Bovid Species as We Thought?" *Systematic Biology*, January. doi:10.1093/sysbio/syt004.

IISE. 2011. "Retro SOS 2000-2009: A Decade of Species Discovery in Review." Tempe, AZ. http://species.asu.edu/SOS.

Isaac, Nick J, James Mallet, and Georgina M Mace. 2004. "Taxonomic Inflation: Its Influence on Macroecology and Conservation." *Trends in Ecology & Evolution* 19 (9): 464–69. doi:10.1016/j.tree.2004.06.004.

Jach, M A. 2000. "International Code of Zoological Nomenclature, 4th Edition" 25 (2): 273–82. http://dx.doi.org/10.1046/j.1365-3113.2000.252107.x.

Jetz, Walter, Jana M. McPherson, and Robert P. Guralnick. 2012. "Integrating Biodiversity Distribution Knowledge: Toward a Global Map of Life." *Trends in Ecology & Evolution* 27 (3): 151–59. doi:10.1016/j.tree.2011.09.007.

Johnson, Ned K, J V Remsen Jr, and Carla Cicero. 1999. "S26.1: Resolution of the Debate over Species Concepts in Ornithology: A New Comprehensive Biologic Species Concept." In *Proc 22 Int. Omithol. Congr.*, edited by N J Adams and R H Slotow, 1470–82. Johannesburg: BirdLife South Africa. http://www.internationalornithology.org/proceedings/Proc22 IOC/Symposium/S26/S26.1.htm.

Jones, Owen R., Andy Purvis, Eligiusz Baumgart, and Donald L. J. Quicke. 2009. "Using Taxonomic Revision Data to Estimate the Geographic and Taxonomic Distribution of Undescribed Species Richness in the Braconidae (Hymenoptera: Ichneumonoidea)." *Insect Conservation and Diversity* 2 (3). Blackwell Publishing Ltd: 204–12. doi:10.1111/j.1752-4598.2009.00057.x.

Joppa, Lucas N, David L Roberts, and Stuart L Pimm. 2011. "The Population Ecology and Social Behaviour of Taxonomists." *Trends in Ecology & Evolution* 26 (11): 551–53. doi:10.1016/j.tree.2011.07.010.

Kennedy, J, R Hyam, R Kukla, and T Paterson. 2006. "Standard Data Model Representation for Taxonomic Information." *OMICS: A Journal of Integrative Biology* 10 (2). Mary Ann Liebert, Inc. 2 Madison Avenue Larchmont, NY 10538 USA: 220–30. doi:10.1089/omi.2006.10.220.

Kennedy, Jessie B, Robert Kukla, and Trevor Paterson. 2005. "Scientific Names Are Ambiguous as Identifiers for Biological Taxa : Their Context and Definition Are Required for Accurate Data Integration," 80–95.

Kociolek, J. P., and David M Williams. 2015. "How to Define a Diatom Genus? Notes on the Creation and Recognition of Taxa, and a Call for Revisionary Studies of Diatoms." *Acta Botanica Croatica* 74 (2). doi:10.1515/botcro-2015-0018.

Kociolek, J P. 1996. "Comment: Taxonomic Instability and the Creation of Naviculadicta Lange-Bertalot in Lange-Bertalot & Moser, a New Catch-All Genus of Diatoms." *Diatom Research* 11 (1). Taylor & Francis Group: 219–22. doi:10.1080/0269249X.1996.9705373.

Lapage, SP, PHA Sneath, EF Lessel, VBD Skerman, HPR Seeliger, and WA Clark. 1992. *International Code of Nomenclature of Bacteria: Bacteriological Code, 1990 Revision*. ASM Press. http://www.ncbi.nlm.nih.gov/pubmed/21089234.

Laurenne, Nina, Jouni Tuominen, Hannu Saarenmaa, and Eero Hyvönen. 2014. "Making Species Checklists Understandable to Machines - a Shift from Relational Databases to Ontologies." *Journal of Biomedical Semantics* 5 (1): 40. doi:10.1186/2041-1480-5-40.

Lepage, Denis. 2017. "Avibase." Edited by Denis Lepage. Bird Studies Canada. http://avibase.bsc-eoc.org/.

Lepage, Denis, Gaurav Vaidya, and Robert Guralnick. 2014. "Avibase – a Database System for Managing and Organizing Taxonomic Concepts." *ZooKeys* 420 (June): 117–35. doi:10.3897/zookeys.420.7089.

Lim, Gwynne S., Michael Balke, and Rudolf Meier. 2012. "Determining Species Boundaries in a World Full of Rarity: Singletons, Species Delimitation Methods." *Systematic Biology* 61 (1). Oxford University Press: 165–69. doi:10.1093/sysbio/syr030.

Linnaeus, Carl. 1753. *Species Plantarum, Exhibentes Plantas Rite Cognitas Ad Genera Relatas, Cum Differentiis Specificis, Nominibus Trivialibus, Synonymis Selectis, Locis Natalibus, Secundum Systema Sexuale Digestas*. Holmiae (Stockholm): Laurentii Salvii. doi:10.5962/bhl.title.669.

———. 1758. *Systema Naturae per Regna Tria Naturae, Secundum Classes,*

Ordines, Genera, Species, Cum Characteribus, Differentiis, Synonymis, Locis. Tenth. Laurentii Salvii. doi:10.5962/bhl.title.542.

Mann, David G. 2010. "Discovering Diatom Species: Is a Long History of Disagreements about Species-Level Taxonomy Now at an End?" *Plant Ecology and Evolution* 143 (3): 251–64. doi:10.5091/plecevo.2010.405.

Mayr, Ernst. 1942. *Systematics and the Origin of Species, from the Viewpoint of a Zoologist*. Harvard University Press. http://www.hup.harvard.edu/catalog.php?isbn=9780674862500.

McNeill, J, F R Barrie, W R Buck, V Demoulin, W Greuter, D L Hawksworth, P S Herendeen, et al. 2012. *International Code of Nomenclature for Algae, Fungi and Plants (Melbourne Code)*. Koeltz Scientific Books. http://www.iapt-taxon.org/nomen/main.php.

Meier, Rudolf. 2017. "Citation of Taxonomic Publications: The Why, When, What and What Not." *Systematic Entomology* 42 (2). Blackwell Publishing Ltd: 301–4. doi:10.1111/syen.12215.

Meyer, Carsten, Holger Kreft, Robert Guralnick, and Walter Jetz. 2015. "Global Priorities for an Effective Information Basis of Biodiversity Distributions." *Nature Communications* 6 (September). Nature Publishing Group. http://www.nature.com/ncomms/2015/150907/ncomms9221/full/ncomms9221.html.

Miller, Jeremy, Donat Agosti, Lyubomir Penev, Guido Sautter, Teodor Georgiev, Terry Catapano, David Patterson, et al. 2015. "Integrating and Visualizing Primary Data from Prospective and Legacy Taxonomic Literature." *Biodiversity Data Journal* 3 (3): e5063. doi:10.3897/BDJ.3.e5063.

Mora, Camilo, Derek P. Tittensor, Sina Adl, Alastair G. B. Simpson, and Boris Worm. 2011. "How Many Species Are There on Earth and in the Ocean?" Edited by Georgina M. Mace. *PLoS Biology* 9 (8): e1001127+. doi:10.1371/journal.pbio.1001127.

Morse, DR, N Ytow, DM Roberts, and A Sato. 2003. "Comparison of Multiple Taxonomic Hierarchies Using TaxoNote." http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.8673&rep=rep1&type=pdf.

Mozzherin, Dmitry Y., Alexander A. Myltsev, and David J. Patterson. 2017. "'gnparser': A Powerful Parser for Scientific Names Based on Parsing Expression Grammar." *BMC Bioinformatics* 18 (1): 279. doi:10.1186/s12859-017-1663-3.

Mullins, Patricia L., Ricardo Kawada, James P. Balhoff, and Andrew R. Deans. 2012. “A Revision of Evaniscus (Hymenoptera, Evaniidae) Using Ontology-Based Semantic Phenotype Annotation.” *ZooKeys*, no. 223: 1–38. doi:10.3897/zookeys.223.3572.

National Audubon Society. 2017. “Audubon Bird Guide App.” http://www.audubon.org/apps.

Olson, Storrs L. 1987. “On the Extent and Source of Instability in Avian Nomenclature, as Exemplified by North American Birds.” *The Auk* 104 (3): 538–42. doi:10.2307/4087560.

Olson, Storrs L, and James L Reveal. 2009. “Nomenclatural History and a New Name for the Blue-Winged Warbler (Aves: Parulidae).” *The Wilson Journal of Ornithology* 121 (3). The Wilson Ornithological Society: 618–20. doi:10.1676/09-003.1.

Padial, José M, and Ignacio de la Riva. 2006. “Taxonomic Inflation and the Stability of Species Lists: The Perils of Ostrich’s Behavior.” *Systematic Biology* 55 (5): 859–67. doi:10.1080/10635150600815888.

Page, Roderic D M. 2013. “BioNames: Linking Taxonomy, Texts, and Trees.” *PeerJ* 1 (1). PeerJ, Inc: e190. doi:10.7717/peerj.190.

Peet, RK, MT Lee, and MF Boyle. 2012. “Vegetation-Plot Database of the Carolina Vegetation Survey.” *Vegetation Databases for the 21st Century. – Biodiversity & Ecology*, 243–253. doi:10.7809/b-e.00081.

Peters, James Lee, G. W. Cottrell, James C. Greenway, Ernst Mayr, Raymond A. Paynter, and Melvin A. Traylor. 1931. *Check-List of Birds of the World.* Cambridge : Harvard University Press,. doi:10.5962/bhl.title.14581.

Peterson, A. Townsend, Sandra Knapp, Robert Guralnick, Jorge Soberón, and Mark T. Holder. 2010. “The Big Questions for Biodiversity Informatics.” *Systematics and Biodiversity* 8 (2): 159–168. doi:10.1080/14772001003739369.

Pimenta, Bruno V.S., Ulisses Caramaschi, and Carlos Alberto Gonçalves Cruz. 2015. “Synonymy of Crossodactylus Bokermanni Caramaschi & Sazima, 1985 with Crossodactylus Trachystomus (Reinhardt & Lütken, 1862) and Description of a New Species from Mina.” *Zootaxa* 3955 (1): 65. doi:10.11646/zootaxa.3955.1.3.

Pimm, Stuart L, and Lucas N Joppa. 2015. “How Many Plant Species Are There, Where Are They, and at What Rate Are They Going Extinct?” *Annals of the Missouri Botanical Garden* 100 (3). Missouri Botanical Garden: 170–76. http://dx.doi.org/10.3417/2012018.

Pocock, Reginald Innes. 1929. “Tigers.” *Journal of the Bombay Natural History Society* 33. Bombay Natural History Society: 505–41. http://www.biodiversitylibrary.org/part/154233.

Potapova, Marina, and Donald F Charles. 2007. “Diatom Metrics for Monitoring Eutrophication in Rivers of the United States.” *Ecological Indicators* 7 (1). Elsevier Ltd: 48–70. http://linkinghub.elsevier.com/retrieve/pii/S1470160X0500110X.

Queiroz, Kevin De. 2007. “Species Concepts and Species Delimitation.” *Systematic Biology* 56 (6). Oxford University Press: 879–86. doi:10.1080/10635150701701083.

Remsen, David. 2016. “The Use and Limits of Scientific Names in Biological Informatics.” *ZooKeys*, no. 550. Pensoft Publishers: 207–23. doi:10.3897/zookeys.550.9546.

Remsen Jr., J V. 2015. “HBW and BirdLife International Illustrated Checklist of the Birds of the World Volume 1: Non-Passerines.” *Journal of Field Ornithology* 86 (2): 182–87. doi:10.1111/jofo.12102.

Ride, W. D. L., H. G. Cogger, C. Dupuis, O. Kraus, A. Minelli, F. C. Thompson, and P. K. Tubbs, eds. 1999. *International Code of Zoological Nomenclature*. International Trust for Zoological Nomenclature. http://www.worldcat.org/isbn/9780853010067.

Ridgway, Robert. 1923. “A Plea for Caution in Use of Trinomials.” *The Auk* 40: 375–76.

Ridgway, Robert, and Herbert Friedmann. 1901. *The Birds of North and Middle America : A Descriptive Catalogue of the Higher Groups, Genera, Species, and Subspecies of Birds Known to Occur in North America, from the Arctic Lands to the Isthmus of Panama, the West Indies and Other Island*. Washington : Govt. Print. Off.,. doi:10.5962/bhl.title.54021.

Rising, James D, and Frederick W Schueler. 1972. “How Stable Is Binominal Nomencature?” *Systematic Zoology* 21 (4): 438+. doi:10.2307/2412436.

Sangster, George. 2009. “Increasing Numbers of Bird Species Result from Taxonomic Progress, Not Taxonomic Inflation.” *Proceedings of the Royal Society B: Biological Sciences* 276 (1670): 3185–91. doi:10.1098/rspb.2009.0582.

———. 2014. “The Application of Species Criteria in Avian Taxonomy and Its Implications for the Debate over Species Concepts.” *Biological Reviews of the Cambridge Philosophical Society* 89 (1): 199–214. doi:10.1111/brv.12051.

Sangster, George, and Jolanda A Luksenburg. 2015. "Declining Rates of Species Described per Taxonomist: Slowdown of Progress or a Side-Effect of Improved Quality in Taxonomy?" *Systematic Biology* 64 (1). Department of Bioinformatics and Genetics, Swedish Museum of Natural History, P.O. Box 50007, SE-104 05 Stockholm, Sweden; Department of Zoology, Stockholm University, SE-106 91 Stockholm, Sweden; and Department of Environmental Science and Policy, George: Oxford University Press: 144–51. http://dx.doi.org/10.1093/sysbio/syu069.

Schulenberg, Thomas S, and Marshall J Iliff. 2014. "Updating the eBird/Clements Checklist 6th Edition." *Birds.cornell.edu*. http://www.birds.cornell.edu/clementschecklist/about/methods/.

Shen, Tsung-Jen, Anne Chao, and Chih-Feng Lin. 2003. "Predicting the Number of New Species in Further Taxonomic Sampling." *Ecology* 84 (3). Ecological Society of America: 798–804. doi:10.1890/0012-9658(2003)084[0798:PTNONS]2.0.CO;2.

Sibley, David. 2012. "Name Changes of Birds in the 2012 AOU Supplement," July. http://www.sibleyguides.com/2012/07/name-changes-of-birds-in-the-2012-aou-supplement/.

Sluys, Ronald. 2013. "The Unappreciated, Fundamentally Analytical Nature of Taxonomy and the Implications for the Inventory of Biodiversity." *Biodiversity and Conservation*, 1–11. doi:10.1007/s10531-013-0472-x.

Stan Development Team. 2017. "RStan: The R Interface to Stan." http://mc-stan.org/interfaces/rstan.

Steege, Hans ter, Rens W. Vaessen, Dairon Cárdenas-López, Daniel Sabatier, Alexandre Antonelli, Sylvia Mota de Oliveira, Nigel C. A. Pitman, et al. 2016. "The Discovery of the Amazonian Tree Flora with an Updated Checklist of All Known Tree Taxa." *Scientific Reports* 6: 29549. doi:10.1038/srep29549.

Stone, Witmer. 1935. "Some Aspects of the Subspecies Question." *The Auk* 52 (1): 31–39. doi:10.2307/4077105.

Stone, Witmer, Harry C Oberholser, Jonathan Dwight, T S Palmer, and Charles W Richmond. 1923. "Eighteenth Supplement to the American Ornithologists' Union Check-List of North American Birds." *The Auk* 40: 513–25. https://sora.unm.edu/node/12912.

Swainson, William, and John Richardson. 1831. "Part Second, The Birds." In *Fauna Boreali-Americana, Or, The Zoology of the Northern Parts of British America: Containing Descriptions of the Objects of Natural History Collected on the Late Northern Land Expeditions under Command of Captain Sir John Franklin,*

R.N. London: John Murray. doi:10.5962/bhl.title.63874.

Swarth, Harry S. 1931. "The Tyranny of the Trinomial." *The Condor* 33 (4): 160–62. doi:10.2307/1363682.

Swick, Nate. 2016. "2016 AOU Check-List Proposals, Part 1," January. http://blog.aba.org/2016/01/2016-aou-check-list-proposals-part-1.html.

Szczepocka, Ewelina, Bogusław Szulc, Katarzyna Szulc, Barbara Rakowska, and Joanna Żelazna-Wieczorek. 2014. "Diatom Indices in the Biological Assessment of the Water Quality Based on the Example of a Small Lowland River." *Oceanological and Hydrobiological Studies* 43 (3). Versita: 265–73. doi:10.2478/s13545-014-0141-z.

Tancoigne, Elise, Cyprien Bole, Anne Sigogneau, and Alain Dubois. 2011. "Insights from Zootaxa on Potential Trends in Zoological Taxonomic Activity." *Frontiers in Zoology* 8 (1). BioMed Central Ltd: 5. doi:10.1186/1742-9994-8-5.

Tancoigne, Elise, and Alain Dubois. 2013. "Taxonomy: No Decline, but Inertia." *Cladistics* 29 (5): 567–570. doi:10.1111/cla.12019.

Taxonomic Names and Concepts Interest Group. 2006. "Taxon Concept Transfer Schema, Version 1.01." http://www.tdwg.org/standards/117/.

The American Ornithologists' Union. 2017. "Committee on Classification and Nomenclature (North & Middle America): Operating Procedures." http://www.aou.org/committees/nacc/.

"The National Water-Quality Assessment Program—Science to Policy and Management." 2010. *Water.usgs.gov*. http://water.usgs.gov/nawqa/xrel.pdf.

Thomer, A., G. Vaidya, R. Guralnick, D. Bloom, and L. Russell. 2012. "From Documents to Datasets: A Mediawiki-Based Method of Annotating and Extracting Species Observations in Century-Old Field Notebooks." *ZooKeys* 209. doi:10.3897/zookeys.209.3247.

Tuominen, Jouni, and Nina Laurenne. 2011. "Taxon Meta-Ontology TaxMeOn." http://schema.onki.fi/taxmeon/.

U.S. Geological Survey. 2017a. "BioData - Aquatic Bioassessment Data for the Nation." doi:10.5066/F77W698B.

———. 2017b. "Integrated Taxonomic Information System (ITIS)." https://www.itis.gov/.

Uetz, Peter. 2016. "The Reptile Database Turns 20." *Herpetological Review* 47 (2): 330–34.

Weakley, Alan S. 2015. *Flora of the Southern and Mid-Atlantic States*. http://www.herbarium.unc.edu/flora.htm.

Wetmore, Alexander, Herbert Friedmann, Frederick C Lincoln, Alden H Miller, James L Peters, Adriaan J van Rossem, Josselyn Van Tyne, and John T Zimmer. 1944. "Nineteenth Supplement to the American Ornithologists' Union Check-List of North American Birds." *The Auk* 61 (July): 441–64. https://sora.unm.edu/node/18734.

Wieczorek, John, David Bloom, Robert Guralnick, Stan Blum, Markus Döring, Renato Giovanni, Tim Robertson, and David Vieglais. 2012. "Darwin Core: An Evolving Community-Developed Biodiversity Data Standard." *PLoS ONE* 7 (1): e29715+. doi:10.1371/journal.pone.0029715.

Williams, D, and G. Reid. 2006. "Large and Species Rich Taxa." In *Reconstructing the Tree of Life: Taxonomy and Systematics of Species Rich Taxa.*, 305–22. doi:10.1201/9781420009538.ch19.

Wortley, Alexandra H, and Robert W Scotland. 2004. "Synonymy, Sampling and Seed Plant Numbers." *Taxon* 53 (2): 478–80. doi:10.2307/4135625.

Ytow, Nozomi. 2016. "Taxonaut: An Application Software for Comparative Display of Multiple Taxonomies with a Use Case of GBIF Species API." *Biodiversity Data Journal* 4 (4). Faculty of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Ibaraki 305-8572, Japan.: Pensoft Publishers: e9787. doi:10.3897/BDJ.4.e9787.

"Zoological Record." 2017. Accessed May 8. http://wokinfo.com/products_tools/specialized/zr/.