Title
Contingent Consensus: Documentary Control in Biodiversity Classifications

Author
Montoya, Robert D.

Publication Date
2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Contingent Consensus:

Documentary Control in Biodiversity Classifications

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Information Studies

by

Robert Delgado Montoya

2017

ABSTRACT OF THE DISSERTATION

Contingent Consensus:

Documentary Control in Biodiversity Classifications

by

Robert Delgado Montoya

Doctor of Philosophy in Information Studies

University of California, Los Angeles, 2017

Professor Johanna R. Ducker, Chair

In order to gain a better sense of the globe's biodiversity there have been concerted efforts within the biodiversity community to aggregate dispersed databases to facilitate universal access to information. Central to these systems are nomenclatural and taxonomic mechanisms that validate, organize, and collocate data using established standards and classifications. This dissertation is about the identification, naming, control of, and access to, this cache of biodiversity knowledge, and the common *information, documentation*, and *classification* problems that materialize as part of this process. Invoking theories articulated in Information Studies, I examine how *documentary control* functions within the biodiversity environment, defined as it is by *contingent* concepts and documents, and how these disciplinary conditions negotiate this tension through classification structures. In particular, *composite taxonomies* are examined as authoritative access-oriented classifications designed predominantly to aggregate multiple biodiversity taxonomies under one management classification to facilitate efficient data

communication. Such composite structures are situated in contradistinction to traditional, descriptive-based taxonomies, primarily designed to argue a hypothesis-driven position about how organisms are related. As *constructed* knowledge organization systems, biological classifications make implicit epistemological and ontological claims about biological facts, yet these attributes are often overlooked in the practice of interfacing with these systems. Given the increased prominence of these databases within scientific and professional communities, this dissertation asks what *kind* of knowledge these composite taxonomies instantiate and represent, and how successful they are in serving a consensus-based taxonomic purpose. Taking a critical Information Studies approach, these issues are explored by deeply analyzing the Catalogue of Life, a prominent composite taxonomic schema, invoking documentary, historical, and qualitative methodologies. This project critiques and illustrates the radiant effects of composite taxonomies in biodiversity networks and their multiple uses in professional and scientific practice. This manuscript argues all knowledge organization systems—biological and otherwise—are constructs of cultural and historical circumstances, manufactured artifacts of certain spatiotemporal positions. My goal is to show how other disciplines can inform the literature, theories, and work within Information Studies to rethink our problems anew. As I see it, the question is no longer whether our classification systems can attain true representational capacities, it is more about how we are going to acknowledge their constructedness and harness their contingencies for the most situated social benefit.

The dissertation of Robert Delgado Montoya is approved.

Geoffrey Bowker

Christopher M. Kelty

Jonathan Furner

Johanna R. Drucker, Committee Chair

University of California, Los Angeles

2017

This work is dedicated to my mother, Elena Montoya, who worked unfathomably hard as a single mother to raise five children. Despite working all day, every day (and many nights), you always found the time to attend even the most minor student teacher conferences. You are inspiring and the reason I continue to push myself to be a better human being. And to my uncle, Alfred "Baca" Delgado, whose life was taken away far-too-soon. Your curiosity and avid pursuit of knowledge were unique and contagious. Confidence in my own identity began with you.

A dissertation is a long, sometimes frustrating process, and I could not have pushed through without the support of my partner, Steve Barnthouse. Steve, all my love to you for being there at the most difficult moments of this process. This manuscript is complete because you believed that it could be.

My deep and sincere thanks to my dissertation committee: Johanna Drucker, Jonathan Furner, Geof Bowker, and Chris Kelty. If I can accomplish half as much as each of you in my career I would count myself accomplished. Thank you for supporting me and being stalwart intellectual models. A special thanks to Johanna: I can say nothing more than that my gratitude is endless. I wish also to thank Ronald Day, Joseph Tennis, Gregory Leazer, Anne Gilliland, Christine Borgman, Jens-Erik Mai, Safiya Noble, Michelle Caswell, Amelia Acker, Bryan Heidorn, and David Ribes for their advice and support.

Of course, this manuscript owes much to my friends and colleagues who helped me think through these issues, or perhaps just as importantly, helped me forget about the dissertation during these years of data collection and analysis, writing, and editing: David Myers, Julie VanWinkle, Rahul Subramanian, Raj Mithal, Marika Cifor, Seth Erickson, Patricia Garcia, Mario H. Ramirez, Angel Diaz, Jess Deshayes, and Roderic Crooks. I would also be remiss without thanking my former colleagues at UCLA Library Special Collections: Jeffrey Rankin,

Lucinda Newsome, Victoria Steele, Genie Guerard, Thomas Hyry, Aislinn Sotelo, Octavio Olvera, Josh Fiala, Amy Wong, Cesar Reyes, Simon Elliott, Jane Carpenter, Jain Fletcher, Brandon Barton, Virginia Steel, Susan Parker, and Sharon Farb. I will always cherish those fine moments working together.

The support I received from the individuals behind the Catalogue of Life and the broader taxonomic community has been absolutely essential to the success of this project. The narrative I present is possible only because of this cooperation. Even in moments when direct citation to the Catalogue of Life is not warranted, every section owes a debt of gratitude to the education in taxonomy and bioinformatics they provided. In particular, I thank Yury Roskov, Thomas Orrell, David Remsen, Mike Ruggiero, Peter Schalk, Thierry Bourgoin, Paul Kirk, Timothy Utteridge, Alan Paton, Nicolas Bailly, Tim Robertson, Doug Yanega, Vincent Smith, Matthew Woodburn, René Dekker, and Jeroen Snijders.

And finally, a shout-out to the libraries, archives, and museums that made this project possible: UCLA Library; Smithsonian Museum of Natural History; Integrated Taxonomic Information System; Naturalis Biodiversity Center; Global Biodiversity Information Facility; Natural History Museum, London; The Linnean Society; Royal Botanic Gardens, Kew; Oxford Museum of Natural History; Special Collections, Department of Zoology, University of Cambridge; American Museum of Natural History; and The Bancroft Library. Last, but not least: the Los Angeles Law Library in Downtown L.A., where much of this manuscript was written. May we never lose focus in supporting the institutions that protect our cultural memory, and provide safe spaces for those that need it the most.

Table of Contents

List of Figures

List of Abbreviations

BBC	Bliss Bibliographic Classification
BHL	Biodiversity Heritage Library
BoL	Barcode of Life
BoLD	Barcode of Life Database
BOLI	Barcode of Life Initiative
CABI	Centre for Agriculture and Biosciences International
CBD	Convention on Biological Diversity
CBOL	Consortium for the Barcode of Life
CoL	Catalogue of Life
CERN	Conseil Européen pour la Recherche Nucléaire/ European Organization for Nuclear Research
DDC	Dewey Decimal Classification
EoL	Encyclopedia of Life
FRBR	Functional Requirement Bibliographic Records
GBIF	Global Biodiversity Information Facility
GDI	General Definition for Information
GGN	Global Genome Initiative
GGBN	Global Genome Biodiversity Network
GIS	Geographic Information System
GNA	Global Names Architecture
GNI	Global Names Index
GNUB	Global Names Usage Bank
GSD	Global Species Database
GTI	Global Taxonomic Initiative
GUID	Globally Unique Identifier
ICNAFP	International Code of Nomenclature for Algae, Fungi, and Plants
ICZN	International Code of Zoological Nomenclature *Also*: Internal Commission on Zoological Nomenclature
IPNI	International Plant Names Index
ICT	Information and Communication Technologies
I/R	Information Retrieval
I/S	Information Studies
ITIS	Integrated Taxonomic Information System (originally Interagency Taxonomic Information System
ITPG	Information Technology Program Group, formerly part of the now defunct Biometrics and Computing Section of the Natural History Museum, London
KO	Knowledge Organization
LCC	Library of Congress Classification System
NBI	Naturalis Biodiversity Institute
NCBI	National Center for Biotechnology Information
NHML	Natural History Museum, London
NMNH	Smithsonian National Museum of Natural History
NSF	National Science Foundation
NYBGIH	New York Botanical Garden Index Herbariorum Database

OCR	Optical Character Recognition (Software)
ORCHID	Open Researcher and Contributor Identification
OTU	Operational Taxonomic Unit
POWO	Plants of the World Online Portal
STS	Science, Technology, and Society
TDWG	Biodiversity Information Standards/Taxonomic Databases Working Group
TNU	Taxon Name Usage
TCS	Taxon Concept Schema
WoRMS	World Register of Marine Species

The research for this dissertation was funded by grants and fellowships from the UCLA Department of Information Studies; Bernard and Martin Breslauer Professor of Bibliography Fund; UCLA Graduate Division, National Science Foundation (Science, Technology, and Society Dissertation Improvement Grant, 1556062); Smithsonian National Museum of Natural History; Litwin Books (Award for Ongoing Dissertation Research in the Philosophy of Information); and Beta Phi Mu International Library & Information Studies Honor Society (Eugene Garfield Doctoral Dissertation Fellowship).

VITA (*Selected*)

I. EDUCATION

Master of Library and Information Science (M.L.I.S.)
University of California, Los Angeles, 2015
History of the Book, Print and Visual Culture Concentration

M.F.A. in Creative Writing (Poetry)
Antioch University, Los Angeles, 2008

B.A in English (American Literature and Culture)
University of California, Los Angeles, 2003
Minor in Anthropology (Biological Emphasis)

II. PUBLICATIONS/CONFERENCE PROCEEDINGS

Montoya, R. D. (2017) "Boundary Objects/Boundary Staff: Supporting Digital Scholarship in Academic Libraries." *The Journal of Academic Librarianship*, 43(3), pp. 216-223.

Montoya, R.D., Erickson, S.R. (2017) "Anachronism in global information systems: the cases of Catalogue of Life and Unicode." *iConference 2017 Proceedings* (Wuhan, China, March 22-25, 2017).

Montoya, R. D. (2016). "Advocating for Sustainability: Scaling-Down Library Digital Infrastructure," *Journal of Library Administration*, 56(5), pp. 603-620.

Cifor, M., Montoya, R.D. and Ramirez, M.H. (2016). "Developing an Undergraduate Information Studies Curriculum in Support of Social Justice." *iConference 2016 Proceedings* (Philadelphia, PA, March 20-23, 2016).

Montoya, R. D. (2016). "A Classification of Digital Emergence: A Critical Approach to the Production of Digital Objects in Special Collections," *Canadian Journal for Academic Librarianship*, 1(1), 42-59.

Miller, K, Montoya, R.D. (2013). "Teaching and Learning Los Angeles through Engagement with UCLA Library Special Collections," *Urban Library Journal*, 19(1).

III. PRESENTATIONS AND PANEL ORGANIZATION

Leazer, G., Montoya R.D. (2017). "Limits of Infrastructure." Panel organized for Society for Social Studies of Science Annual Meeting (Boston, MA, August 30-September 2, 2016).

Montoya, R.D. (2017). "Toward and Eco-Documentary Justice in Information Studies." Paper presented at the Libraries and Archives in the Anthropocene ("LAAC17") Colloquium (New York, NY, May 13-14, 2017).

Montoya, R.D., Erickson, S.R. (2017). "Anachronism in global information systems: the cases of Catalogue of Life and Unicode." Paper presented at iConference 2017 (Wuhan, China, March 22-25, 2017).

Montoya, R.D. (2016). "On the Functionality of Taxonomic Documents." Paper presented at the Annual Meeting of the Document Academy (Denton, Texas, September 30-October 1, 2016).

Bowker, G., Montoya R.D. (2016). "New Topologies of Scientific Practice." Panel organized for Society for Social Studies of Science Annual Meeting (Barcelona, Spain, August 31-September 3, 2016).

Montoya, R.D. (2016). "Database Projection: Repositioning Knowledge in Biodiversity Taxonomic Databases." Paper presented at the Society for Social Studies of Science Annual Meeting (Barcelona, Spain, August 31-Septemebr 3, 2016).

Montoya, R.D. (2016). "Articulating Composite Taxonomies: Knowledge Organization and The Catalogue of Life." Paper presented at the Catalogue of Life Symposium (Heraklion, Crete, Greece, April 14, 2016).

IV. TEACHING (UNIVERSITY OF CALIFORNIA, LOS ANGELES)

Teaching Associate. General Education Clusters, First Year Initiatives. Los Angeles: The Cluster (Fall 2016, Winter 2017, Spring 2017).

Adjunct Instructor/Faculty, California Rare Books School (CalRBS). Course: Better Teaching with Rare Materials (Summer 2016).

Teaching Fellow/Instructor of Record. Information Studies 30: Internet and Society (Summer 2016).

V. OTHER PROFESSIONAL EMPLOYMENT

Head of Public Services, UCLA Library Special Collections Public Services, University of California, Los Angeles; August 2014-September 2015.

Operations Manager and Communication Officer, UCLA Library Special Collections Public Services, University of California, Los Angeles; January 2011-July 2014.

Reader Services Coordinator, UCLA Library Special Collections Public Services, University of California, Los Angeles; June 2007-December 2010.

Assistant Director, Historical Society of Southern California; April 2006-June 2007.

Chapter 1: Information Studies, Biodiversity, and Ecosystemic Approaches to Knowledge Organization

> *In [Knowledge Organization] we make implicit statements about knowledge of concepts, acts (such as representation), entities, and systems. In doing so, we create knowledge, and our epistemic stance dictates what kind of knowledge that is.*
>
> *—Joseph Tennis*
> *"Epistemology, Theory, and Methodology in Knowledge Organization: Toward a Classification, Metatheory, and Research Framework" (2008, p. 103)*

> *The individual scientist's view of the world is shaped in cultural-historical and disciplinary contexts which influence their criteria of, among other things, classification.*
>
> *—Birger Hjørland, Eric Scerri, John Dupré*
> *"Forum: The Philosophy of Classification" (2011, p. 14)*

> *The purpose of a classification is to provide a simplified reference system that is biologically sound and widely useful. It should be compatible with the phylogeny, but it cannot serve its central simplifying purpose unless it leaves out some of the fine detail about relationships that are essential for some phylogenetic purposes. One can use a phylogeny as a basis for making a classification, but one cannot logically deduce a fully detailed phylogeny from a classification. Nor is a phylogeny sufficient to give a classification. A phylogeny and a classification must be congruent (i.e. not contradictory) but they are different ways of abstracting from and representing biological relationships.*
>
> *—Thomas Cavalier-Smith*
> *A Revised Six-Kingdom System of Life (1998, pp. 212–213)*

> *In scientific work the aim is not so much to discover properties of kinds, or to define essence, as to select a particular property of interest for some scientific purpose, and to determine the conditions of its appearance. The purpose is not by study of the property to prove the nature of the thing that has it, but to fix or control the circumstance under which a thing possesses a property necessary to prove something else … The theory of properties is not so far removed from causal explanation, for to say, "this thing has the property of such and such" is merely a way of saying "this thing under certain conditions behaves in such and such way." … Whereas logical division would merely co-ordinate species, "as being all alike members of one differentiation of a common element," the arrangement of books must take the form of a serial order, since books are constructed so as to be placed in a row. Had they been products of nature, existing in every conceivable shape, no serial arrangement could have intruded into their classification, and they would have been classified in some other pattern, less obliterative of distinctions.*
>
> *—A. Broadfield*
> *The Philosophy of Classification* (1946, pp. 93–95)

Identifying a Global Biodiversity Taxonomic Impediment

In June of 1992, a meeting of nations gathered at the United Nations Conference on Environment and Development in Rio di Janeiro, Brazil (collectively called the "Earth Summit"), to discuss some of the most pressing issues facing the sustainability of biological life on the planet, "leading to the adoption of Agenda 21, a wide-ranging blueprint for action to achieve sustainable development worldwide" (United Nations, 1997, 2017). Topics of the

Summit included the lack of access to potable water, the increasing use of fossil fuels and the concomitant production of toxic waste, and the decline of the globe's biodiversity due, in particular, to the detrimental influence of human activities (Department of Public Information, 1997). A critical international treaty arising from this Summit, the "Convention on Biological Diversity" (CBD), documented the critical need for coordinated scientific information-exchange infrastructures in order to better understand and reverse the globe's diminishing biological diversity (2016). A core focus of the CBD was the acknowledgement that biodiversity was much more than merely the identification of "plants, animals and micro organisms and their ecosystems" (2017c), but also included the radiant influence this information has upon global populations (*all* biological populations, including, but not limited, to humans), as well as the research practices of professional scientists that engaged in this work on a daily basis.

The historical importance of this document as it pertains to worldwide biodiversity has been relatively significant, for it served as a motivating and *pro forma* agreement upon which numerous biodiversity and ecological initiatives could build support and endorsement for their respective projects. On a local level, the CBD has influenced the drafting and implementation of laws and policies that govern a number of domains pertaining to biodiversity issues (Kate, 2002). The CBD has arisen as a watershed moment in biodiversity studies, and a catalyst and authority for *coordinating* geographically local knowledge within openly accessible, global intellectual infrastructures of information exchange.

In particular, Article 7 of the CBD explicitly acknowledged the importance of maintaining and organizing "by any mechanism data, derived from" ("Convention on Biological Diversity (full text)," 1992, p. 5) the identification and monitoring of biological diversity.

Building on this acknowledgement, during the 1998 Conference of the Parties,[1] participants established the Global Taxonomic Initiative (GTI) to rectify what participants called the "taxonomic impediment" toward the successful implementation of the CBD organizational and access goals (Convention on Biological Diversity, 2017b; Hopkins & Freckleton, 2002). The impediment identifies a "shortage of taxonomic expertise, taxonomic collections, field guides, and other identification aids, as well as to the difficulty in assessing existing taxonomic information" (Convention on Biological Diversity, 2003, p. vii). According to the CBD, The GTI marks "the first time in history that taxonomy has had recognition at such a high level in international policy" (2003). Given the CBD's articulated concerns with the fragmented nature of biodiversity knowledge, the GTI was meant to articulate clear steps by which authoritative online platforms could potentially collocate regional taxonomic information by strengthening "networks for regional cooperation" (2003, p. 1). Locally-specific knowledge had grown too fragmented, too many researchers were "[hoarding their] data," and with limited funding to go around, centralizing databases seemed to be one mechanism by which scientists could sustain a long view approach (Ribes & Finholt, 2009) to describing the world's organisms (Thomas, 2009). In response to the Global Taxonomic Initiative, large-scale federated platforms began to gain operational steam, aggregating taxonomic and descriptive data with the goal of unifying geographically specific caches of biodiversity knowledge (Bowker, 2008, p. 120; Waterton, Ellis, & Wynne, 2013, p. 108).

Despite the existence of initiatives such as the GTI, global biodiversity, according to scientific literature, continues to decline (Bellard, Bertelsmeier, Leadley, Thuiller, &

[1] "The Conference of the Parties is the governing body of the Convention, and advances implementation of the Convention through the decisions it takes at its periodic meetings" (Convention on Biological Diversity, 2017a).

Courchamp, 2012; Butchart et al., 2010).[2] A more robust understanding of the biodiversity of the planet is a continually pressing agenda, essential to efforts seeking to fully document the extent (and, increasingly, extinct) number of biological species. Scientist's ability to study and understand the scope of any given ecological issue, however, rests on the scientific community's capacity to name, document, and classify, the collective knowledge regarding biological taxa (their circumscription, interrelationships, and ecology) for easy access, sharing, and utilization in research, policy-making, and conservation efforts. Uniform, publicly accessible, up-to-date biodiversity knowledge is essential if scientists are to pool their efforts and engage in scholarly conversation. The assumption underlying this coordinated approach is that such global knowledge provides the foundation for larger-scale approaches to biodiversity problems.

The problematics and machinations of coordinating and producing such biodiversity taxonomic infrastructure, however, are numerous and complex, particularly if we look more closely at the specific technological databases that have been created to aggregate this knowledge. In the last twenty years—spurred on by, if not created as a direct result of, the Convention's articulated aims and directives—new federated digital initiatives such as the Global Biodiversity Information Facility (GBIF) (2015), the Catalogue of Life (Species 2000, 2015b), Encyclopedia of Life ("Encyclopedia of Life: Global access to knowledge about life on Earth," 2015), and the Barcode of Life (2015; Waterton et al., 2013), have taken on the

[2] Of course, there is much debate about how we define and categorize aggregate biodiversity knowledge, and how such approaches and mechanisms lead to the *mis*conception that biodiversity is decreasing in aggregate (Dornelas et al., 2014; Vellend et al., 2013). Assessment of biodiversity loss depends upon many factors, including spatial and temporal research emphases (McGill, Dornelas, Gotelli, & Magurran, 2015). Results differ between regional and local examinations, and "paleontological data show that life is surprisingly resilient" (2015, p. 104). In general, certain charismatic species are given more attention than others (Bowker, 2008, p. 146), forcing developments regarding the growing number of extant and proliferating bacterial species in the background of these debates (Locey & Lennon, 2016). Yet, despite the differences of opinion on the matter, "the majority of models indicate alarming consequences for biodiversity, with the worst-case scenarios leading to extinction rates that would qualify as the sixth mass extinction in the history of the earth" (Bellard, Bertelsmeier, Leadley, Thuiller, & Courchamp, 2012). Given the importance of species variety to the ecological health of *all* geographies, widespread scientific opinion is that biodiversity loss is a pressing issue that requires coordinated efforts to address.

management of worldwide biodiversity data toward the end of universal access and standardization of information, and represent some of the world's most robust aggregators of biological *description* and *control*. These information systems are collectively used in scientific research to direct global biodiversity initiatives supported by governmental and non-governmental organizations in the development of "effective policies," and to make decisions "regarding land management, health, climate change and biodiversity conservation" (Jetz, McPherson, & Guralnick, 2012, p. 151).

This dissertation is about the identification, naming, control of, and access to, this cache of biodiversity knowledge and the common *information, documentation*, and *classification* problems that materialize as part of this process. Given that biodiversity scientists and informaticians perform these activities in particular contexts, I will also necessarily look to how this *control* is articulated and maintained within certain limited organizational and social frameworks.

Disciplinary Framing

"The world is full of writings," wrote Patrick Wilson, "How can the valuable be kept from oblivion? How can [one] be sure of finding, in the great mass of writings, good and bad, pedestrian and extraordinary, the writings that would be of value to [them]?" (1968, p. 1). Long has it been the tradition in Information Studies[3] to focus on the bibliographic tradition in relation to classification systems. But *information fields* are concerned with far more than traditional bibliographic documents. The information disciplines, or "disciplines of the cultural record," as articulated by Marcia Bates, are orthogonal in nature, meaning "they deal with every traditional

[3] A note is warranted related to my use of "Information Studies" (I/S) throughout this document. For the purposes of simplicity and brevity, I will use I/S as the umbrella term that also encompasses information science as well as archives.

subject matter" by asking particular questions about the "collection, organization, retrieval, and presentation of information in various contexts" (2007, pp. 1–2). The information disciplines deal with the myriad of *documentation* forms produced within these numerous disciplines, broadly conceived in this dissertation to refer to any signifying set of objects, markings, and signs, that act as evidence for some kind of process, activity, object, or phenomena of interest. Whether our focus of attention is a library, a museum, or a database, the primary issues of concern from an *information* point-of-view are how these units of documentation relate to particular fields of human activity. As Bates remarks, "the universe of living things throws off documentary products, which then form the universe of documentation" (2007, p. 7). The world is, indeed, full of writings, but it is also full of information, data, and documents, that we identify as significant in particular contexts for particular purposes—which we then classify and organize so as to provide access to our collective public knowledge (Wilson, 1977). This organization, however, is more than pragmatic; it is also an emergent knowledge in-and-of itself, framing, as it does, documents into discreet presupposed ontological and classificatory categories, which are then doubly presupposed to relate to one another based on various likenesses and differences (Broadfield, 1946, Chapter 1).

This manuscript, then, is very much a work about the core concepts of *information*, *documentation*, and *classification* as they are articulated and theorized in Information Studies. In particular, I am primarily concerned with the analysis of *biodiversity classifications* on a global scale. Aside from my longstanding general interest in the subject, biodiversity taxonomies are interesting and applicable to our discipline because they force new perspectives onto the literatures, theories, and approaches, central to Information Studies. It is too often the case that, when asked to describe my research to colleagues in Information Studies, the reactionary

question is, *Why biodiversity studies?* My response to this is, *Why anything?* But it is more than that. What concerns me slightly about this general inquiry is that topics and concepts such as information, documentation, classification, and taxonomy are either conceptualized primarily in relation to bibliographic systems, popular classification schema such as the Library of Congress Classification (LCC) system, Dewey Decimal Classification (DDC), and other classification schema; or as web based ontological schema and semantic languages. In other words, the wide scale default assumption is usually that classifications should organize *all* knowledge into universal systems. Divergent, extra-disciplinary, and discipline specific forms of classification are often considered an *atypical* course of study in the field of I/S.[4] Of course, this is not a categorically negative assumption, for much of this dissertation is meant to be *in service* to these, and other, general approaches to organizing knowledge. Moreover, the intent is to deepen our understanding of niche classificatory approaches so as to more adequately produce flexible and extensible systems that can more adequately represent divergent points of view and epistemological stances.

The problem (if that is not too strong a word), as I see it, is that not enough individuals at the current moment are looking at the "generation of theories, principles and methods that emphasize both the cultural and historical specificity of classification practices and their emancipatory function" (Furner, 2013a). If they were, then initial reactions to my topic of analysis would be less confounding to some. It is useful, I think, to exit our spaces of disciplinary comfort to find surprises in these specificities. My argument here is that we need *more* of this

[4] Certainly there are exceptions to this rule in the Information Studies literature, especially within the subfield of Knowledge Organization. Scholars, and in particular Birger Hjørland, have long advocated the utility for looking at domain specific (Asundi, 2012; Deokattey, Neelameghan, & Kumar, 2010; Hjørland & Albrechtsen, 1995; Hjorland & Hartel, 2003) knowledge organizing systems. Hjørland's case study of the classification of psychology (1998) is a good example of such as approach, as have other articles in the journal, *Knowledge Organization*, looking at the production of scientific classification in various contexts (Blake, 2011; Gnoli, 2006; Hjørland & Nicolaisen, 2003; Hjorland, Birger, Scerri, & Dupre, 2011; Marco & Navarro, 1993).

work, for the situatedness, contextually-specific, and historically-contingent attributes of these specific systems can inform, broaden, and render more pluralistic, our understanding of classification and knowledge organization in general. Such work, I believe, is vital to longevity of our discipline, and keeps theories within I/S relevant and knowledgeable of current knowledge organizing concerns. Secondly (and admittedly most anecdotally), I also see fewer scholars than I would hope engaging with those authors they suppose to be essential reading on the syllabi for our core Information Studies courses. While there are surely names that I am omitting, I see the work of Patrick Wilson (Wilson, 1968, 1977, 1983), A. Broadfield (1946), Henry Bliss (Bliss, 1929, 1933), Elaine Svenonius (Svenonius, 2004, 2009), Seymour Lubetzky (1969), among many others, as some of these individuals.[5] This dissertation is also my modest attempt at re-engaging with these scholars on *contemporary* terms, hoping to restate their significance in relation to problems of current concern.

The fundamental questions I ask here regard how we can think about extending the concepts of Information Studies outward to think about classificatory domains that otherwise have received very little notice in our discipline, or at least that have not received much in-depth notice in quite some time. I hope this project will produce a balanced view of the strengths and limitations of such biodiversity approaches to Knowledge Organization (KO) in the biodiversity sciences, as well as how such approaches can somehow inform practices within Information Studies. Additionally, I wish to bridge the classifying and standardization activities that occur within biodiversity studies with similar literatures and activities within Information Studies. As Ronald E. Day states, "Universal bibliographical classifications and descriptions followed the example of zoological taxonomy and classification in the century before them" (2014, p. 39).

[5] Again, certain authors—who I cite extensively—*do* look to these intellectual predecessors, such as Jonathan Furner, Joseph Tennis, Jens-Erik Mai, among many others, and I would like to think of my work as being in conversation with this long line of individuals.

Seen in this light, this project seeks to return back to these roots, to reengage I/S scholarship in the organizing endeavors and practices of the natural sciences. Taking a close look at the practices of biodiversity scientists in relation to practices in Information Studies isn't an altogether strange juxtaposition, for as David Hull states, "as most people view taxonomists, they are more librarians than scientists and just as loveable…collectors and classifiers were [and, I would argue, still are] the ones who had sufficient knowledge to appreciate the true diversity of life" (1988, p. 81).

For one, comparing the theories and practices of biodiversity classification—concerned with biological entities—with the organization of documents in Information Studies has highlighted a rather important *difference* between how these approaches conceptualize the *object* of classificatory concern in relation to the *system* in which it is embedded. In Information Studies, our aim is to organize a set of books or documents, their classificatory and ontological specificity within a system of organization is anchored *only* by the artificial boundaries and suppositions imposed by the classifier. Books and documents are not created *out of* a *natural* system or ecology that makes their classification self-evident. Such artificial classificatory concepts are what drove Hope Olson, in her influential text, *The Power To Name: Locating the Limits of Subject Representation in Libraries* (2002), to push against the "fundamental presuppositions" on which our information practices rest (in Olson's case, she was focusing on subject representation within library systems). If there is a defining system that constrains the ontological specificity of texts, many have stated that *disciplines* fit the bill. Many of the most widely-used classification system schedules, including the LCC, DDC, and the Bliss Bibliographic Classification (BBC), take disciplines (or disciplinary areas of study) as the primary classificatory unit through which documents gain their cognitive and institutional

authority (Wilson, 1983, Chapter 4). Disciplines change and are redefined over time, however, while schedules change ever so slowly, if at all, to account for these ongoing developments. Within these socially-constructed disciplinary systems the terms by which the description of documents function is idiosyncratic and can take many forms since there is no presupposed *natural* organization of disciplines upon which classifiers can say, 'this is the definite system,' and by which they can compare or contrast classificatory outcomes.

In the biodiversity realm, however, there certainly is an extent to which 'the real' plays a fundamental role in *how* and *why* we classify things the way we do. After all, "from the beginning, one of the chief goals of science, possibly the chief goal of science, has been to discover classes of phenomena that are lawfully related—classes commonly termed natural kinds" (D. L. Hull, 1988, p. 78). Biodiversity classifications are unique in that they engage with ostensibly natural occurring objects (species and taxa) that can *empirically* be examined and assessed for subsequent coordination in classifications. The biological objects can be assessed in many ways, using any number of traits: morphological, genetic, ecological, etc. But as Johanna Drucker indicated in a private communication, "there is no equivalent to DNA in a book, rather, it is subject to the interpretive activity of description." For biodiversity classifiers there is a definite system of natural objects that can serve as a ground for comparison and contrast between different approaches. However, as the chapters of this dissertation will illustrate, even with such a notion of the 'real' grounding the act of classification, there is no still no presupposed way in which this system can be *translated* into representational classifications. In fact, biological classifications are *arguments* and *hypothesis* for how we can understand these natural classes according to current scientific research. Classifications—even those empirically based—are models, not mirrors. Each scientist will have a different take on what natural kinds exist and how

these kinds are related to one another. The end result is that even classifications based on natural phenomena are subjective and presuppose a number of socially constructed presuppositions.[6] How we interpret a natural object as it relates to the natural phenomena in which is embedded will change over time; science is not static, our understanding of the world changes as new information and understandings are built. This distinction between documentary and biological systems is a key one to keep in mind as you read this manuscript. The contrast between these approaches, and how they construct and verify the existence of a classifiable object, is a space that can tell us a great deal about the subjective and representational qualities of knowledge organizing schema more broadly speaking.

But this project is much more than a demonstration of how two related scholarly domains can engage in productive discussions, it is also a way in which we can articulate the *importance* of Information Studies theories, literatures, and methodologies to domains that otherwise remain blind to the robust theoretical and methodological work the discipline has to offer. Our discipline has something to add, for as Jonathan Furner has stated, "there are several scholarly communities other than information studies that do require a separate concept of information, but that those communities have good reason to look to information studies for help. Any approach to conceptualizing information that downplays the contributions of LIS— i.e., information without information studies—is needlessly impoverished, not least on account of the range of ontological possibilities that it misses" (2014). In addition to *information*, I would add

[6] As I will also discuss in chapter four of this manuscript, there are some scholars in the field of Knowledge Organization that subscribe to a phenomena-based approach to classification (Claudio Gnoli & Riccardo Ridi, 2014; Gnoli, 2009; Gnoli & Poli, 2004; Rick Szostak, 2008). Such approaches assume that phenomena can be structured according to "integrative levels" (Gnoli & Poli, 2004) that see reality as organized into strata that then consist of varying levels of complexity (2004, p. 152). Reality—and documents—then, can be organized and classified according to this complexity by the application of descriptive facets. If the study of biodiversity taxonomy tells us anything, it is that phenomena are *not* static and interpretively unrestricted, and that such an approach does not and cannot produce a universal knowledge system that conforms to a uniform and consistent model of reality any more than one that arises from a disciplinary approach to organization.

documentation, classification and knowledge organization literature as a worthwhile set of writings that have a great deal of *value* to other scholarly communities.

Broadening Traditional Knowledge Organization in Information Studies

In this manuscript I take a relatively broad view of the practice of KO, one that extends beyond the organization of documents and the design of systems that provide adequate representations of those documents for mere *retrieval* (Hjørland, 2008, p. 86). Rather, I embrace a definition of KO that embodies the organization of *all concepts* as the derivative mechanism by which we control the "production and dissemination of 'knowledge'" (2008, p. 86). Given the role taxonomic systems play in the organization of biodiversity (and biological knowledge more generally), I see these infrastructures as an equal topic of consideration for Information Studies.

KO and classification theory is a robust subdomain in I/S research. Joseph Tennis defines KO as "the field of scholarship concerned with the design, study and critique of the process of organizing and representing documents that society see as worthy of preserving" (2008, p. 103). Building on this definition, Hjørland adds "works and concepts" to the list of KO concerns, as well as a more "narrow meaning" of KO, that includes those practical activities familiar to [Library and Information Science] audiences, such as bibliographic description, indexing, and the classification of documents in "memory institutions" (Hjørland, 2008, p. 86). KO literature can broadly be seen as encompassing a number of other narrow concerns, such as the development of controlled vocabularies for description; the building of classification systems; the articulation of entity relationships within KO systems; the development of taxonomies; the effective *use* and "goodness" (Furner, 2009) of KO systems; the philosophy and theory of KO; and the social practices that support such infrastructures. This list is not exhaustive, but it serves to indicate the breadth of this sub-domain, and how its concerns are suitable to examining how

biodiversity taxonomies are structurally producing a certain *argument* about how biological organisms relate to and function within our global database ecosystem.

KO in this dissertation is understood to embrace the broader, more inclusive definition advocated by Hjørland above, and by doing so, embraces the practices taking place in the production of biodiversity taxonomies. Hjørland's move to include not only "document representations," but also "works and concepts" (2008, p. 86) is a significant one. On the one hand, we can understand the *concept* in this case to refer to concepts-as-subjects, such as the articulation of the subject of documents (and other derivatives) in cataloging, indexing, or bibliographic classification systems (as in document *x* is *about* the subjects *y* and *z*). On the other, I also take this to mean the representation of *concepts* more broadly and generally understood. Metaphysically speaking, concepts as understood in this project can be as broad as the "kinds (a.k.a. categories, classes, sorts) and individuals" articulated by Jonathan Furner in "Type-Token Theory and Bibliometrics" (2016b). Such kinds could include the concept of species and/or taxa for any given set of biological organisms that, as described above, are anchored to an external reality.

As classification systems have evolved, they bring with them the vestiges of epistemological approaches specific to their circumstances, bounded by classificatory "inclusions and exclusions" (Olson, 2002, p. 6) that are evidence of particular domain subjectivities. Information Studies scholars Birger Hjørland and Jenna Hartel help us understand the extent to which the classificatory and epistemological commitments of such domains find their way into the knowledge organizing infrastructures that define how we partition information objects of interest, and biological classification exemplifies this phenomenon. They state, "It is critical to understand that domains are dynamic. Knowledge production and knowledge organization are

not just about the addition of new elements into pre-established classification. As knowledge develops and evolves, the view of structures of the world and the relations between different concepts changes symbiotically" (2003, p. 244). The dynamism proposed here is of core importance in this manuscript, for the divergent (and multiple) interpretive frameworks imposed upon biological classifications are a fundamental hindrance to providing a unified structure for information coordination.

Threads of Inquiry and Composite Taxonomies

This work examines how biodiversity knowledge is collected, represented, organized, and delivered within domain-specific *composite taxonomies* by examining how such taxonomies *instantiate* particular notions of what constitutes information entities, documents, relationships, representations, and knowledge itself. Composite taxonomies are understood in this project as *taxonomic infrastructures that aggregate multiple taxonomies into one, authoritative infrastructure, in an effort to unify biodiversity data from multiple sources*. More specifically, this dissertation takes as its case study the knowledge organization practices of the Catalogue of Life as a pivot point and primary object of study, to examine how this taxonomy articulates and expresses biodiversity information. The Catalogue is an "authoritative" biodiversity schema and framework that strives to serve two functions in the biodiversity database ecology: (1) to provide "a single integrated species checklist," as well as (2) "a taxonomic hierarchy" around which scattered biodiversity data can be appended onto and organized, and through which it can subsequently be accessed (2015a). Thus, this is not only a project about classification, it is also a project about the representational building blocks that make these classifications function with a certain amount of purchase in the biodiversity community. This work will also necessarily describe how species concepts are documented in these systems, as well as how nomenclature

(as text-strings) serve as tokens around which species-specific data are collocated. There are many layers of a classification that must be peeled-back to understand how it is they *work* as both a container for produced knowledge, as well as an architecture for a new kind of emergent interpretative framework.

To examine the Catalogue's composite taxonomy, as well as its attendant layers and components, I follow two main lines of inquiry simultaneously, both of which have direct bearing on the practices and theoretical foundations of Information Studies. The research questions driving this dissertation include,

1. How can biodiversity taxonomic practices be brought into conversation with the theories about information, documents, and concept representation within Information Studies?
2. How might we understand *composite taxonomies* as information systems designed to *both* represent biological knowledge *and* coordinate efficient data communication?

Two distinct points I think that are essential to examine: one, that there is epistemological and ontological work that happens quite separately from the space of taxonomic production. This research brings to the fore the *constructedness* and *artificiality* of classification systems in general. And of course scholars working in the domain of knowledge organization and classification know this to be true—after all, the first short section in Elaine Svononius's, *The Intellectual Foundation of Information Organization* (2009), states that intellectual foundation upon which a system of organization rests consists of the ideology, formalized processes, research generalizations, and research foci of the discipline in question (2009, p. 1). Exposing this foundation is part of my goal here. Knowledge organization is as unique within communities—scientific or otherwise—as the theories, methods, and objects of analysis, that define their particular domain of study. The biodiversity taxonomic profession illustrates and

amplifies this fact. Each taxonomy, generated by every scientist, is produced under a certain set of intellectual conditions: assumptions about what comprises a species or taxon, as well as engrained suppositions about how these concepts can then be related in various ways based on any number of morphological, genetic, or ecological traits. The resultant taxonomic representations, classifications, taxonomies, trees of life, etc., are shaped by the minute and seemingly infinite intellectual and methodological assumptions of those that create them. Systems of any kind, and no less taxonomic classifications, are contingent historical reconciliations, based on current and present knowledge-sets with an equal footing in the laboratories "of the past" (Rheinberger, 2010, pp. 89–90). If you pick a taxonomy—any taxonomy—it is a network of knowledge that represents years and perhaps decades of layered and accumulated information and research.

Despite this unavoidable and complicated reality, the Catalogue has taken it upon itself, to the best of their efforts, to comingle these diverse and multiple taxonomic constructions into one unified space. This is not apolitical work, nor universally recognized as effective by the taxonomic community. Real conflicts arise as practical and pragmatic approaches to data collection and collocation are positioned in tension with the hermeneutic and hypothesis-driven work of scientific taxonomic production. Yet, the Catalogue's stance is that *information must be shared* in order for biodiversity knowledge to reach its full research impact and potential, and in order for such facilitation to take place, standards need to be implemented, even if taxonomies must be manipulated as they are ingested into a global taxonomic framework. And so, a composite taxonomy can help Information Studies better understand the virtues and downsides of this integrative approach.[7] Concerned as the Information Studies community is with pluralistic

[7] I certainly do not want to give the impression that my critiques indicate, in any way, that the Catalogue (or any other system) is doing something *incorrectly*. Systems take time to produce and while the long view of infrastructure

approaches to classification (Mai, 2011; Szostak, 2015) and the representation of diverse voices and fluid ontologies in and for our information systems (Seddon & Srinivasan, 2014; Srinivasan, Boast, Furner, & Becvar, 2009; Srinivasan & Huang, 2005; Srinivasan, Pepe, & Rodriguez, 2009), spaces such as those inhabited by the Catalogue can be incredibly instructive toward these just ends.

But internal taxonomic comingling in the Catalogue is just one valence of the story; the Catalogue is *also* integrated into *other* systems as core organizational data architecture. Once the Catalogue is compiled, it can (and is) subsequently embedded into a network of other biodiversity systems, thereby amplifying its effect across the discipline of biodiversity and taxonomic studies. Contemporary database taxonomies are now the *main* source of taxonomic knowledge, structured in such a way for reuse and manipulation for a number of constituents (Hodkinson, 2011; Parr, Lee, Campbell, & Bederson, 2004; Watson, Lyal, & Pendry, 2015, Chapters 2, 9). As such, representation of knowledge within these systems, and the constitution of its classificatory knowledge should be closely scrutinized for what it renders invisible in its infrastructure, including the assumptions about taxon definitions, nomenclatural control

---

may seem a simple task to implement from the outside, the short view from those that need to coordinate the structure know all-too-well that even the simplest actions require negotiation and time to conceptualize, plan, and process, as well as to gain support from within the biological and taxonomic community. My goal is to think about what might be working or not working from an information management standpoint—in service to the documentary, library, and museum knowledge systems that Information Studies is primarily concerned with—and to identify how epistemological choices and approaches in these taxonomies play a role in what synergies and conflicts I was able to identify in my research and fieldwork. Secondly, I do not purport to have any radical new insights into the process of nomenclature and taxonomy for biodiversity scientists in specific; while I have grown acquainted with the field during these three years of research on this subject, I am far from a biologist or taxonomist by training. This said, what I *have* learned in this process is that taxonomists and biodiversity informaticians are an incredibly self-aware and meta-analytic community that are always willing to examine their own practices and discuss particular insights into, and blind spots within, their discipline and practices. As biodiversity systems become the primary method of communication, nomenclature and taxonomic groups are becoming seriously focused on how to mediate scientific practice within the limitation of networked and computational systems. This renewed organizational and disciplinary focus proved a fertile environment for research and study.

processes, local taxonomic practices and interpretations, as well as the historical evolution of all of these intellectual activities.

The narrative of this dissertation, then, begins at the buildings blocks of these taxonomic systems (information, documentation, databases, type specimens, species literature), then moves outward to nomenclatural control, discusses the composite taxonomy of the Catalogue of Life, then finally articulates how this taxonomy is utilized as a knowledge base in multiple ways.

Method of Examination

The question then becomes how I managed to gain access to the distributed data necessary to situate and analyze the Catalogue's systemic effects. In the earliest phase of this project, I had a lengthy conversation with Science, Technology, and Society (STS) scholar, David Ribes, now Associate Professor in Human Centered Design and Engineering at the University of Washington, about my preliminary thoughts on how I was to approach biodiversity global taxonomic coordination. During this discussion he asked what seems to be a very basic question, but one that had a huge impact on the framing of my overall theoretical approach: Is your project for people in the discipline of STS studies or the discipline of Information Studies? The truth is, of course, I hope that this work has some relevance in each of these disciplines, but certainly decisions had to be made, not only theoretically, but also methodologically about how I was to approach my object of analysis. It became clear, quite quickly, that my approach to biodiversity taxonomic coordination was to emphasize *information* problems, supported by an examination of *socio-technical* issues as they relate to the political and cultural issues that shape and are shaped by distributed technologies. Of course, this does not mean that I ignore the social and cultural valences, and production of biodiversity knowledge and systems, in these arenas, or the ways in which information objects alter these social arrangements. Nor does it mean that STS

studies does not ever take information approaches—one need look no further than the work on standards, classification, and information infrastructure to see these connections are alive and strong (Bowker, 2008; Bowker & Star, 1999; Bowker, Timmermans, Clarke, & Balka, 2015). It does mean, however, that this manuscript favors what I consider to be core Information Studies literature, while drawing from STS literature related to biodiversity studies and databases generously as appropriate in the conversation.

As will be expanded upon below—and progressively through the course of this manuscript—the Catalogue of Life is a highly integrative, fragmented, and hybrid entity. To say the Catalogue is the object of study for this project is to invoke a number of institutions, policies, research groups, and technical infrastructures that are involved in its maintenance. Given this distributed framework, there are a number of key individuals, sites, and archives that have been integral to this project's success. Studying how and why such a composite taxonomy is necessary in the biodiversity field required an equally fragmented approach to studying its composition. The project, thus, makes use of multiple modes of analysis. First, and primarily, it takes a historical and documentary analytic (often critical) approach as its core methodology, tracing how concepts within biodiversity studies can be seen in conversation with the literature and philosophy of Information Studies. This theoretical approach is augmented by a series of in-person and video interviews with scientists and informaticians from all over the globe, both within and without the Catalogue's staff and administration. In order to gain a balanced sense of the Catalogue's functions and influence, speaking to individuals unassociated to the Catalogue was significant here. It is important to note that the production of a system like the Catalogue of Life is a relatively new initiative in the biodiversity sciences, therefore understanding the nitty-gritty, so to speak, of *how* it works technically and informationally required a good deal of travel,

conversation, and intent listening. Research required flying over 34,500 miles, visiting four foreign countries, nine field study sites, perusing seventy-five archival collections, and interviewing twenty-five individuals (some on numerous occasions). The general architecture and emphases of the dissertation was dictated by these travels, conversations, and data. What was most important to the matter at hand rose to the top, what was least applicable was, unfortunately, omitted from this narrative. This is not the end of the story, only the first phase of a much longer research trajectory.

Mapping the Integrative Landscape: The iLife Consortium

> *In the early days [the Global Biodiversity Information Facility (GBIF)] considered a broad scope of options. The organisation considered implementing species pages; tracking, digitizing and indexing literature; building a global taxonomy - a catalogue of life; they considered museums and specimen data; they talked about observational data. The early founders of GBIF originally considered implementing work programmes covering all of those. It was decided that they didn't have the [ability] to do everything, or the resources. The focus was put on the occurrence [data] ... I understand that decision opened the doors for Encyclopedia of Life to take ownership of the species aspects; BHL for the literature; increased investment in the Catalogue of Life... It is interesting though, that originally it was seen ... as a global biodiversity project ... and since taxonomy binds/links all of this together, and that much of the science is intrinsically linked to literature, and to the specimens we need to be working very closely together across those aspects.*
>
> *—Tim Robertson, Head of Informatics, Global Biodiversity Information Facility (2016)*

The production of a global taxonomic database like the Catalogue of Life is one thing, but it is quite another task to take that data and coordinate it across the globe in multiple online platforms that, together, comprise a much larger biodiversity ecosystem. As Tim Robertson's statement above makes clear, these multiple entities have been conceptualized as comprising a larger data-sharing consortium from the very beginning of the process. Biodiversity data is, at its heart, data about global phenomena, but that does not mean that said data is shareable globally. Paul Edwards emphasizes this distinction and shift from "making global data" to "making [that] data global" in his examination of climate data (2010, Chapters 8, 10). In order for the Catalogue to coordinate taxonomic knowledge on a global level, this dissertation will occasionally refer to other biodiversity platforms within the larger integrated landscape—together, all of these

constitute what I am calling the iLife consortium (See Figure 1, below). I'll now sketch out the contours of this iLife space, beginning with the Catalogue of Life, and then provide the briefest snapshot of other platforms that will play a part in the forthcoming analysis; emphasizing, in particular, how they intersect with the Catalogue of Life functionally within this online ecology.[8]

Figure 1. The Life Partnership. Source: Thomas Orrell and Peter Schalk (2016).

The Catalogue of Life: A consensus and composite global taxonomy.

Though the Catalogue will be thoroughly examined in this manuscript as the main focus of analysis, a brief overview of its architecture is warranted to serve as a starting point for the discussion. The Catalogue is a federated database that has two core functions significant to this study: (1) it has the most comprehensive listing of all known existing species of the planet, and, (2) it produces a consensus-based management composite classification to organize the many taxonomies it ingests as part of the nomenclature collocation process.

[8] In the interest of brevity, I have not listed all iLife participants as depicted in Figure 1 as part of this description—I have chosen to focus only on those platforms that play a significant role in my analysis.

The first aspect of the Catalogue that is pertinent to our discussion is its species checklist. Species names are valuable biodiversity tools. Names serve as the unifying agent in a sea of documentation, biological evidence, and produced scientific literature. Species checklists aim to be a "universal and complete" reference that identifies what species exist in a particular area, and without this information, " we can not sustainably use, explore, monitor, manage and protect biodiversity resources" (Species 2000, 2015a). Species checklists can be compiled as part of local ecological surveys (Kalamath Bird Observatory:, 2017) or for national purposes, and function most effectively if they are "integrated, coordinated and disseminated from a single platform" (Hamer, Victor, & Smith, 2012, p. 1). Species on lists such as the International Union for Conservation of Nature Red List (International Union for Conservation of Nature, 2017) are embedded in policy actions based on their inclusion on the list. Similarly, the Catalogue's taxonomic structure, embedded within other information systems, and browsed for species information, is functional because of the relationships that the document presents for each included piece of information. The Catalogue of Life is attempting to create a checklist for all known species that exist on the planet.[9] As of November 2016 it contained more than 1.6 million species, populated by over 150 individual databases from around the globe. [10] An "Annual Checklist" is published once yearly to solidify taxon groups and to allow for comparative studies of its development, while an ongoing dynamic (monthly) version allows for up-to-date, yet less easily citable, reference throughout the year. Established taxonomic editors control all data

[9] The Catalogue of Life has recently begun integrating fossil data into their checklist, but this phase is in its preliminary stages and thus is far from complete. Species counts as reported by the Catalogue also do not include fossil records (Species 2000, 2017b).

[10] This is the species goal articulated by the Catalogue of Life, though numbers differ widely as to how many species *actually* exist on the planet (Eng, 2016; Hug et al., 2016; Zimmer, 2016). David Hill's Open Tree of Life Project projects 2.3 million species (Hinchliff et al., 2015; "'Tree of life' for 2.3 million species released," 2015), while a recent study estimated over one trillion microbial species alone on the planet (Locey & Lennon, 2016). This dissertation will use the Catalogue's estimated count as a matter of convenience and consistency.

accepted into the system. Synonymy fields are built into the Catalogue to allow for variable terminology for species—an issue that remains common in most classification systems, including the Catalogue.

The second part of the Catalogue that we will examine is its consensus management classification. The management classification is "reviewed by experts, not merely aggregated by computers," (Species 2000, 2015b) which is especially important in this analysis because other systems (such as GBIF) take a more computationally mediated approach.[11] The Catalogue uses one, authoritative "taxonomic classification (also called a hierarchy) for management purposes," and uses this classification *above* the node of attachment of each database. *Beneath* this node it uses the classification provided by the [Global Species Database]" (Species 2000, 2014). While there have been numerous attempts to aggregate taxa and names in the biodiversity world, "the main difference of the Catalogue of Life in Frank [Bisby's] mind," the founder of Species 2000 and one of the principle coordinators of the Catalogue of Life, "was that we are doing this through [the] selection of authoritative taxonomic treatments on a global scale for each particular taxon ..." (Roskov, 2016a). So, while the Catalogue of Life is unique in many respects—namely through it's highly curated structure—it is not the only management taxonomic structure. The Catalogue differs from these other management structures in that it collates existing taxonomies to build the base of its infrastructure. Unlike similar systems, such as the National Center for Biotechnology Information (NCBI) Taxonomy Project, which curates a similar kind of management taxonomy based on gene sequencing additions to GenBank, the Catalogue absorbs a number of existing taxonomic infrastructures and curates a singular taxonomy in an effort to organize all biodiversity in one infrastructure (Federhen, 2003). Thus, the NCBI backbone is a

[11] Again, I provide no particular opinion as to which approach may be effective or not, I merely identify the benefits and downsides of these approaches during the course of this discussion.

singular taxonomy that progressively adds individual species as sequences are added to their database, while the Catalogue is comprised of multiple, whole and unified taxonomies, within one structure. It is this quality that makes the Catalogue a *composite taxonomy* that unifies multiple taxonomic trees with varying "epistemic stances" (Tennis, 2008, p. 103) under one umbrella infrastructure. Secondly, the Catalogue is unique in its scope and influence, given that the taxonomy is deeply integrated with the iLife consortium in general and is used as the hierarchical backbone for a number of highly influential and heavily used database infrastructures.

The Catalogue drew a great deal of attention within the scientific community when it formed and became accessible in 2001, most quite optimistic and excited about the prospect of an aggregative, authoritative taxonomic system (Reichhardt, 1999; Bisby, Shimura, Ruggiero, Edwards, & Haeuser, 2002; Cachuela-Palacio, 2006; Gewin, 2002). The Catalogue is comprised of two previously independent entities that merged in June 2001 (Bisby et al., 2002): the Species 2000 organization (2015b), which covers species across the globe (with an original emphasis on European species), and the Integrated Taxonomic Information System (2016), a biodiversity database-sharing partnership of North American organizations. The primary motivator for this merger and the creation of the Catalogue of Life vision was the late Frank Bisby, who believed that biodiversity data "globalization and interoperability" (Bisby, 2000) were essential to motivate future courses and directions in biodiversity science. Species 2000 is the current "legal body for the global Catalogue of Life programme, holding its Intellectual Property Rights, copyright, domain names, access licenses, Memoranda of Understanding (MoU), taking responsibility for continuity between major projects and providing the ongoing governance of the global programme. It is structurally a federation, owned and governed by the participants that

become its members" (Species 2000 Secretariat, 2015a). Based at the Smithsonian Museum of National History, The Integrated Taxonomic Information System (ITIS) is the central, authoritative species checklist for North American species. ITIS "is the result of a partnership of federal agencies formed to satisfy their mutual needs for scientifically credible taxonomic information" (ITIS, 2017a). Based on communication with the Catalogue of Life editor in January 2017, ITIS had contributed about 50% of the Annual Checklist in 2000. Though detailed data from the 2000-2004 datasets is no longer easily available, ITIS comprised 158,884 of the 220,000 core species in the taxonomic Catalogue of Life database in 2005—a full 72% of the total database species count. Comparing these figures with more recent data from the 2016 release, ITIS contributions have increased by only a small margin to 159,821 total species; however, given the Catalogue now totals 1,640,969 species, ITIS now comprises about 9.7% of the total database environment (Species 2000, 2016e).

The subsidiary component datasets that make-up the Catalogue's data set—Global Species Databases (GSD) and Regional Species Databases (RSD)—come from various locations around the world, including databases provided by Kew Royal Botanic Gardens, the World Register of Marine Species (WoRMS) (2017c), Fishbase (2017), and Systema Dipterorum (Pape & Thompson, 2017), which comprise the largest groups of these contributing databases (Species 2000, 2015d). Over time, as more and more GSDs and RSDs are added the Catalogue, species checklist becomes more robust.[12] The Catalogue stands at the center of a multi-tiered

[12] Of course, while the general trend of the Catalogue's core species and GSD count tends to increase over time, decreases occasionally occur as GSDs redact their data from the system (see Appendix B and the species count drop between 2012 and 2013). Comparing GSD data sets for these years, one sees significant changes amounting to a 51,926 species count drop in 2013. Databases added in 2013 included, the Freshwater Animal Diversity Assessment (FADA) Project; Catalogue of Life China; the World Checklist of Freshwater Mollusca; Psocodea Special File (lice taxonomic groups); and the WoRMS Mollusca database. Equally important are those databases that are no longer contained in the list, including AlgaeBase; the WoRMS sea cucumber database; and the Rotifera database (Species 2000, 2016a, 2016b). The significance here is that these fluctuating numbers represent entire taxa that are removed

infrastructure, ingesting subsidiary databases from regional hubs from around the world (Figure 2).

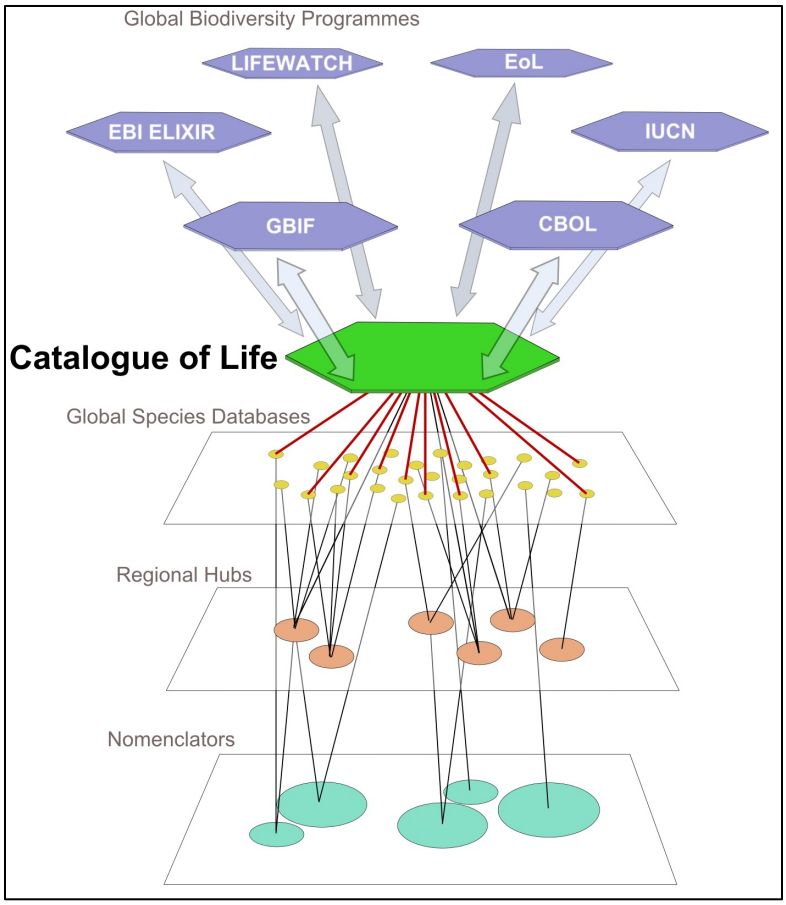


Figure 2. Catalogue of Life Infrastructure Layers (Species 2000, 2015b). Nomenclators exist at the bottom of the infrastructure layer and include all code governed nomenclatural acts, including original name usages (in taxonomic literature), objective synonyms, as well as other name forms (see chapter three for more information). Regional Hubs are regional checklists (RSDs) for a given geographic area. Global species databases (GSD) are those databases that (typically) collect one taxa on a global scale (usually at the genus or family level)—all instances of that genus regardless of geographic boundaries (see chapter four). The Catalogue of Life then aggregates GSDs and RSDs into a master species checklist and composite consensus management taxonomy. The Catalogue is then used as a taxonomic backbone for many other online systems, including GBIF, EoL, and the International Union for Conservation of Nature (IUCN) Red List of Threatened Species (see chapters four and five).

The Secretariat for the Catalogue rotates on a five-year basis, at which point institutions "bid" for hosting privileges. The Catalogue is currently based at the Naturalis Biodiversity Center in Leiden, Netherlands, after a stint at the University of Reading (Evans, 2013). The

---

or added en bloc, creating a cascading effect in numerous domains of biodiversity practice as the Catalogue is updated in the numerous platforms that integrate its taxonomic backbone.

Catalogue's Executive Editor—currently based at the Illinois Natural History Survey, Prairie Research Institute at the University of Illinois, Urbana-Champaign—with the support of a team of editorial experts, collocates the taxonomic and species information, and makes this taxonomic infrastructure freely available as the backbone to a number databases comprising some of the world's most significant aggregators of robust biological description. These partner systems, such as The Encyclopedia of Life, the Global Biodiversity Information Facility, The Barcode of Life Initiative, and the International Union for Conservation of Nature (IUCN) Red List of Threatened Species, focus more on descriptive species content, image data, geographic information system information, and occurrence data. These heavily used online systems are collectively used to direct global biodiversity initiatives supported by governmental and non-governmental organizations. Scientific research is scaffolded and supported by their data. Despite its smaller citation count, the Catalogue is a fundamental, central structural node of ecological infrastructure, serving as the hidden architecture for numerous systems central to the endeavor of biodiversity research (Parr et al., 2012, p. 100, illustration), and as such, its structural, representational, and epistemological attributes should be better understood as a document of biodiversity knowledge and classificatory practice.

Global Biodiversity Information Facility (GBIF).

Amid the iLife consortium GBIF is arguably the most visible, and certainly the largest, aggregator of data points of all participants. GBIF began in 2001, and was founded as a space where biodiversity datasets could easily be published in an open access environment, with the expectation that user groups of all types could explore and download this data to facilitate a more cooperative, globally contextualized biodiversity research (Global Biodiversity and Information Facility, 2013b). The GBIF Secretariat is currently located "at the Natural History Museum in

Copenhagen, Denmark, [and] is charged with developing, executing and reporting on the GBIF work programme" (GBIF, 2013). Similar to the Catalogue, GBIF functions through a distributive structure, with the Secretariat supported by a Governing Board and various standing committees and task groups populated by scientists and professionals around the world. GBIF has a unique stature in the iLife environment as an intergovernmental collaboration, charged with the coordination of a series of global nodes (operational bases) that coordinate the collection and provisioning of data to the global environment—these nodes are often located in a prominent natural history or other biodiversity-related institution within the node country. As of November 2016, there are currently fifty-six governments that have agreed to a non-binding memorandum of understanding, promising to establish a Secretariat that will take the responsibility of maintaining partnerships and coordinating internal institutional network development and contribution of data (GBIF, 2016b, n.d.). Additionally, GBIF is a core source of data for the production of "downstream science"—science that is produced using GBIF as a primary set of core data (T. Robertson, 2016). In 2015 alone, 407 articles were identified as using GBIF-mediated data, and trends show considerable increases since 2005 when only 52 articles were identified as having done so (GBIF Science Committee, 2016).

Given its stature, relatively large staff support, and consistent funding, GBIF plays a significant role in the general discussion of the Catalogue's functionality and future trajectories. In particular, GBIF is one of the largest iLife participants that integrate the Catalogue's management classification into the architecture of their taxonomic backbone (Döring, 2016). GBIF has invested heavily in the coordination of nomenclature in technical environments, and participated in the "NAMES in November" Conference (Global Biodiversity and Information

Facility, 2016b) that is heavily cited later in this dissertation.[13] In chapter five, I describe how GBIF plays a central role in assessing the limitations of the Catalogue's curated taxonomic space, seen especially through GBIF's recent initiatives to enhance their Nub Taxonomy (another term for a backbone taxonomy) via computationally-mediated algorithmic methods (Global Biodiversity and Information Facility, 2016a).

Biodiversity Heritage Library.

The Biodiversity Heritage library (BHL)—an initiative also hosted at the Smithsonian National Museum of Natural History, along with ITIS—is a biodiversity literature repository that coordinates digitization efforts across a number of natural history and other biodiversity scientific institutions. As stated on their website, "the BHL consortium works with the international taxonomic community, rights holders, and other interested parties to ensure that this biodiversity heritage is made available to a global audience through open access principles" (Biodiversity Heritage Library, 2017). Given the intellectual property and copyright issues associated with published items, the majority of the work in BHL is dated pre-1923, and all digital documents are hosted on the webservers of the Internet Archive. BHL uses optical character recognition (OCR) software to identify name-instances that are currently identified and available in the Global Names Architecture.[14] These identified names then allows for the collocation of documents from multiple institutions in one digital space. Mechanisms are currently being put in place to extract names from BHL for immediate inclusion into the GNA, which would provide a workflow mechanism by which newly located species name-tokens can

[13] See chapter three for more information.

[14] See chapter three for a more extensive discussion of the Global Names Architecture. GNA is "a system of web-services which helps people to register, find, index, check and organize biological scientific names and interconnect on-line information about species" (Global Names Architecture, 2017c).

be assessed (and subsequently absorbed) into the Catalogue of Life Plus's nomenclatural workflow.[15] Once BHL identifies particular scientific names in their OCR'd documents, they can then be imported/attached to species-specific webpages on the Encyclopedia of Life.

Encyclopedia of Life.

The Encyclopedia of Life (EoL) is an open-access environment hosted at the Smithsonian NMNH[16] primarily intended to collocate species-specific data from various sources (GBIF and BHL being two of the primary data stores) for display on their one "webpage for every species" (2016) platform. Alongside any given species content page, EoL publishes a series of possible taxonomic arrangements (2017d), of which the Catalogue of Life is one.[17] The curatorial data model of EoL differs significantly from the Catalogue—in EoL's system, a series of curatorial roles are assigned (master, full, and assistant curators), that have varying levels of curatorial responsibility over the veracity of included data, as well as for setting default species display preferences for species that fall within their species expertise, including the default taxonomic arrangement for any given species. According to EoL staff, the primary user base for their services are the general public and K-12 educators—users that want information displayed in a convenient and user-friendly fashion, but that would not otherwise download data for scientific and/or research purposes.

International Commissions on Biological Nomenclature.

Numerous international codes exist to provide guidelines for the application and control of biological species. Prominent codes include, the International Code of Nomenclature for

[15] See also chapter three for information on Catalogue of Life Plus.

[16] Incidentally, ITIS, EoL, and BHL are all located within the same wing of the Smithsonian National Museum of Natural History.

[17] See chapter four for more information about the use of the Catalogue of Life taxonomy on their platform.

algae, fungi, and plants (ICNAFP); the International Code of Zoological Nomenclature (ICZN); the International Code of Nomenclature of Bacteria (ICNB); International Code of Nomenclature for Cultivated Plants (ICNCP); and International Committee on Taxonomy of Viruses (ICTV). Nomenclatural codes govern the global production and control of all names, regardless of geography of language. Only valid scientific names (those correctly published and available in public repertories) are managed through these processes—common names, for example, have no professional relationships with the codes. For the most part, when I discuss the codes of nomenclature from this point forward, I will draw upon the two most significant codes for the Catalogue of Life as well as any of the associated infrastructures that have been discussed: the ICZN (International Commission on Zoological Nomenclature, 1999) and the ICNAFP (International Association for Plant Taxnomy, 2011).

Toward a "CERN" collaboration for biodiversity informatics.

The importance of this wide-scale iLife coordination was reiterated time-and-time again in the meetings I attended and interviews I conducted. The benefits of such collaboration and integration have numerous benefits, particularly at the local level, including the pooling of both fiscal and human resource, and the elimination of work, labor, and technical redundancy that results from the fragmented nature of biodiversity science. Such long-range, higher-level coordinative efforts have been compared to the approach taken by the Conseil Européen pour la Recherche Nucléaire (CERN) initiative (2017), a Geneva-based particle physics laboratory providing a suite of instruments, tools, and complexes, whereby scientists from all over Europe (and the world) can collectively utilize these resources for research purposes. The underlying motivations for a CERN collective were two fold: to capitalize on the increased interest toward "favoring collaborative European bodies like the European Economic Community," and to offset

the reality that "no single European state had either the financial or human resources needed to build the big laboratories that were key to the future of physics" (Galison & Hevly, 1992, p. 81). The same tendencies and problems can be said to apply to taxonomic and biodiversity infrastructure in today's contemporary scientific communities, albeit on a much smaller and less complex technological scale.[18] As taxonomist, marine biologist, and Catalogue of Life Executive Secretary, Peter Schalk, indicated,

> Seeing how [the Life partnership] fits together and how, all together, we are [becoming] a CERN. And that's what we told the [European Commission]—if you want to build a European taxonomic facility with all of these elements in different countries, you're looking at an investment of between 500 million Euros to about one billion Euros. But if you do that and you connect the name component to the molecular component, the gain and yield financially is bigger. And I think for most countries, the running cost of facilities like this will come down. And [then] the focus can be more about science ... rather than tinkering with the engine all the time. In the next 10 years you will have something like CERN. [The] Catalogue as index: a species list and management hierarchy. The question is, how are these functions working together or separately? You need a management hierarchy because if you download a dataset from, say a Diphtheria database, it goes from species to family level to order, etc. The rest is not [there] because the guy who built, he knew where it belongs. You need to put it somewhere, though, to make one system (2016a).

The goal here, not unlike CERN, is to find ways to centralize the data and work functions in some of the biggest natural history museums in Europe, including the Natural History Museum at the University of Copenhagen; the Natural History Museum, London; Museum für Naturkunde, Berlin; French National Museum of Natural History; Senckenberg Natural History Museum (Naturmuseum Senckenberg); and Naturalis Biodiversity Center in Leiden, Netherlands (collectively and anecdotally called the "Big Six"). So much of the *informatics work* in biodiversity science is focused on the technical facilitation and coordination of data in centralized service hubs. This collaboratory-type infrastructure ("Science of Collaboratories (Home)," 2010) could potentially serve to free up scientist's time for primary scientific activities—those activities on and by which a scientist's professional value and work is defined

[18] By "less complex" I mean that biodiversity science has relatively low-level technical requirements in comparison to the infrastructure required of particle physics. While databases require a great deal of memory (increasingly so, as museum and natural history assets are digitized, and as databases proliferate and are continually downloaded and duplicated), the software and hardware requirements for such work are a *considerably* lower-level investment than a hadron collider.

and assessed at an institutional level. As Schalk explained, the six institutions listed above, comprise approximately 120 staff members (of all ranks) working on information and communication technologies (ICT), and thus the merger of these institution's technical efforts frees-up local staff to engage in more traditional scientific activities.[19] The idiosyncratic nature of these multiply-produced biodiversity data caches produces the need for *both* centralized instrumentation (informatics knowledge and computational hosting capabilities) and intellectual taxonomic mediating services to serve as the architecture for said instruments.

New Sharing Networks, Familiar Problems

An important point to note at this juncture, before we embark on a more detailed analysis of the Catalogue, is that the *kinds* of problems experienced in collocating biodiversity data within these technical spaces have long been an issue for biodiversity institutions. One of the major methodological difficulties I encountered in examining a distributed structure like the Catalogue of Life, especially in the early stages of beginning research, was being able to understand the variety of work involved in producing data, and the extent to which, traditionally and historically, this work has been performed in silos within natural history institutions. Geoffrey Bowker has performed considerable research in the biodiversity sciences (2000b, 2000a, 2008; 2015), and throughout this body of work he articulates how the collaborative and aggregative nature of the *work* of biodiversity sciences inherently produces a number of tensions and complications. One essential tension is that each of these institutional silos produces data specific to their purposes, which may or may not conform to the data kinds and structures that exist within other silos. Further, data structures change as technological advancements are integrated

[19] Others have noted that the skills of ICT professionals do not necessarily overlap with those of scientists, so while the freeing-up of staff in the ICT sector of an institution may free up human resources, those resources may not funnel directly into the production of more taxonomic science.

into the practices of performing scientific work—much of the work prior to the 1980's is primarily in paper form and conformed to no global data standard (Figure 3).

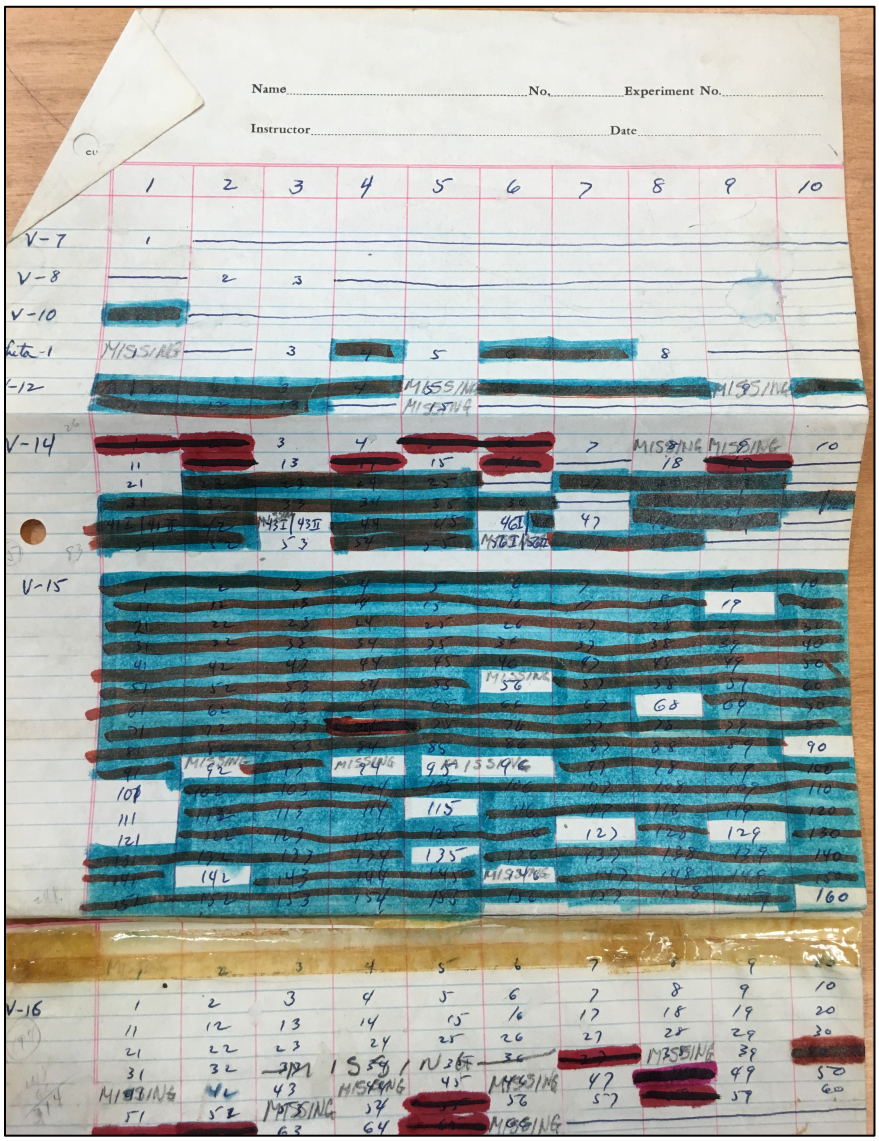


Figure 3. Undated specimen list from Vema (ship) Expedition. American Museum of Natural History Archives *("Specimen list, n.d., color-coded list of specimens," 1955)*. The codes on the left—V-7, V-8, etc.—are specimen numbers that correlate to physical specimens collection on the voyage. Used by permission.

In August of 2016, I spent a good deal of time in the Natural History Museum, London (NHML), which has historically been highly influential and active in the biodiversity informatics community. Serendipitously, I came upon a series of boxes containing the archival documents

produced by the Biometric and Computing section of the NHML (then referred to as the British Museum of Natural History—BM(NH)). The Biometrics and Computing Section was primarily concerned with "inter-departmental compatibility software" (D. H. Thorpe, 1985), much of which contained data that was taxonomic in nature. "The first mention of a coordinated approach to the development of computer (IT) strategy was made in a recommendation (no. 2) of a report of a visiting group to the Biometrics and Computing section in March of 1980" (D. H. Thorpe, 1985). In response to this report, the Biometrics Section formed the Information Technology Program Group (ITPG) in order to identify the various data collecting activities, including the types of software and information organizing systems in use at the NHML; the programming languages being used; word processing packages being implemented; database and taxonomic databases being compiled to maintain the output of research activity (Thompson, 1987), as well as the data formats being collected by both digital and analog means. The end result of this survey could just as well be articulated in the current global biodiversity taxonomic and informatics landscape:

> The current use of computers in the area of collections was unsatisfactory, both from an organizational point of view, and the degree to which requirements are fulfilled. The scientific departments are currently responsible for servicing their own requirements for computing, but this has led to:
>
> 1. Computing section support is only used in a percentage of activities
> 2. Incompatible systems are being used
> 3. In several departments a scientist must spend a proportion of their time looking after computer systems. This is sometimes reluctantly, and is sometimes felt to damage their career prospects.
> 4. Site-license agreements are not used
> 5. Insufficient funding leads to inadequate systems (Information Technology Program Group, 1985).

In the mid-1980's, when this proposal was distributed to NHML administrators, computational capacities were fairly limited, so unified data stores were explicitly ruled out as a possibility. In lieu of such technical collocation, unifying data *standards* across units was proposed, as was the identification of redundant tasks across departments to "reduce duplication of effort" (1985).

By 1989, however, new software models were being proposed and implemented to "provide taxonomists with user-friendly tools for creating and managing data which describe organisms, and to process this into identification keys, classifications, and descriptions" (Thompson, 1989). Noted as well was the importance of nomenclature in the process of data aggregation. ITPG interest then shifted toward the articulation of a computing strategy that could connect these disparate data stores and provide an adequate means of access (without compromising intellectual property or other privacy issues associated with the data, in a pre-internet and web environment) (Thompson, 1988). Four possible options for managing biodiversity data within the NHML were proposed (Figure 4), including a model where (Option 1) all "computing activity was … undertaken on a large main frame or quite of machines; (Option 2) "each department will ultimately have their own system" (as well as being responsible for its own IT strategy); (Option 3) "resources are shared by topic … for example, a single machine would be procured for all taxonomic computing for use by all departments that wish to use the computer in this way"; and (Option 4) "a hybrid between 2 and 3, [where] each department would ultimately obtain their own system…There would be no direct communication between departmental systems but each system would be able to communicate with [the] JANET" United Kingdom research network (1988).[20] The default mode for data management, if agreement on a structure could not reached, was that the museum would treat departments "as five separate museums and each department [would] have to look after its own computing needs" (Information Technology Program Group, 1985, p. 5). Ultimately, in response to this particular initiative, no centralized structure was agreed upon and the siloed approached remained in effect.

[20] The JANET network is "a high-speed network for the UK research and education community" that traces its roots back to the 1970s (Joint Information Systems Committee, 2017).

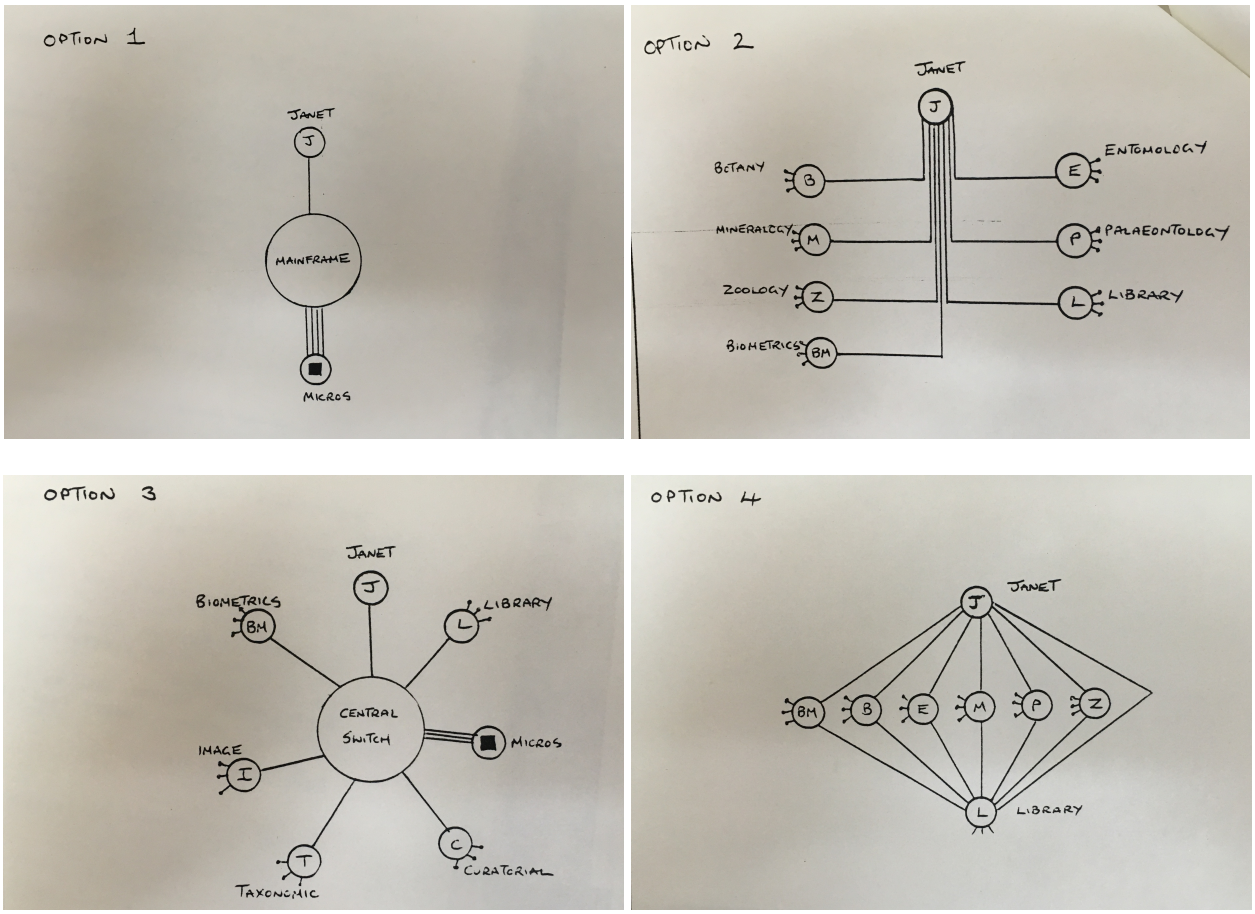


Figure 4. Possible network configurations for data control and sharing at the Natural History Museum, London (Thompson, 1988). Used by permission. © The Trustees of the Natural History Museum, London.

Even with such historical efforts to centralize biodiversity data, the current Informatics team at the NHML continues to deal with incommensurable data created within numerous museum departments. Five data silos that harken back to those identified by the ITPG remain deeply engrained in institutional culture,[21] even taking into consideration the *significant* progress made by the informatics team to centralize data for global sharing (Smith, 2016; Woodburn, 2016).[22] One issue to note—that will be brought up again and again in this manuscript—is the extent to which data is "flattened," filtered, or altered, as it enters the communal data space. The

[21] These department "silos" include zoology, entomology, mineralogy, paleontology, and botany.

[22] The NHML makes its research and collections data available through its online Data Portal (Scott & Smith, 2017), which, via an application program interface (API), contributes data to infrastructures such as GBIF, which then makes it way to other platforms via data sharing protocols. The Data Portal

flattening of data is the selection of local (or departmental) data that only conforms to the standards and capabilities of the centralized database structures.[23] Additionally, the NMNH uses the open-source data management system and portal software, CKAN (2017), which has a particular data structure that only allows the NMNH to have a single table as a dataset. Thus, in CKAN, because you cannot build a relational or maintain data relationships, most data that comes into the Museum's data portal is denormalized, meaning that multi-value fields maintained within Museum departments are being flattened down to conform to this denormalized structure.[24] Further, the Museum's collection management system is about nine separate modules—a core module for the catalogue, a taxonomy module, a collection events module, multi-media module, etc.—and much of this information is not relevant for the data portal or available for public consumption. In the case of the NHML, much of this filtering is performed for good reasons—data may be embargoed while research is ongoing or awaiting publication, species data may identify the location of endangered species, etc.—but this fact does remind us to be constantly aware that the data landscape, as it exists in the iLife consortium, is only a small fragment of all *possible* data types that exist within the local, institutional environments that contribute data to these aggregated structures.

The problems with sharing biodiversity data on a global scale, then, are long familiar to taxonomists. The issues we see at plat in the NHML above are only amplified in a global space, where multiple institutional protocols, practices, and data types must be merged together into a functional network of database fields. What makes the contemporary climate in taxonomy so

[23] Three-dimensional image data, for example, is not included in the NHML's central database, which is then filtered outward into systems such as GBIF.

[24] It was noted by NHML staff that flat, denormalized data files are generally what users request when they need datasets. However, extracting the relationality of multi-value fields back out again for users who wish to maintain that information is incredibly difficult, if not impossible, in the current staffing and technical environment.

exciting is that a critical mass of scientists are starting to see the benefits and importance of such data sharing to both their current research, as well as to the field of taxonomy as a whole (plagued as it is with the "impediment" of diminishing practitioners and expertise). And given this momentum, the Catalogue of Life—with its curated pool of species names and its consensus management taxonomic architecture—stands at the center of this iLife consortium, primed to shape the collocation and access mechanisms of extent species knowledge.

Introduction to the Chapters

This dissertation constitutes my attempt at understanding the kinds of coordination required to create global taxonomic knowledge structures by way of the Catalogue of Life. The chapters take a graded approach to examining this phenomena, beginning with the most basic concepts that populate these databases—species concepts and names—then I progressively expand the scope of my analysis outward to include nomenclatural control; biodiversity taxonomic instruments in general; the consensus-based composite management classification of the Catalogue; and finally ending with an examination of taxonomies as knowledge bases, and the limitations of said instrumentation in the field of biodiversity and within taxonomic practice.

The following chapter, "The Documentary Universe of Biodiversity Databases," examines the fundamental informational units of biodiversity databases and proposes an expanded work-entity conceptual model that can help us understand the production of document entities in database documents like the Catalogue of Life. I begin with Patrick Wilson's notion of the bibliographical universe and progressively broaden his construct outward by including documents, broadly conceived, as an integral part of this new contemporary informational and documentary distributed database space. It seems to me that the fundamental entities that inhabit Wilson's bibliographical universe still hold much relevance—one glance at the Functional

Requirement for Bibliographic Records illustrates this fact—including his concepts of *works*, *texts*, *exemplar*, and *events/objects*, but they can also benefit from some reevaluation and expansion to conform to activities that are performed in online biodiversity documentary spaces. In essence, this chapter looks at the Catalogue as a document-generating document, and attempts to define the universe of documents that it instantiates as a defined biodiversity *text* publisher. In order to accomplish this, I identify the various *kinds* of *entities* defined within the biodiversity database environment, beginning with information, progressing through data, documents, and knowledge. *Documents* are defined as the fundamental entity in this analysis, acting as the representational element for all biodiversity *evidence*. An argument is then made that databases themselves (and biodiversity taxonomy databases in this particular case) are *contingent documents*, defined as they are by continual change, a distributed format, and defined in relation (contingently) to other database documents (and taxonomies). As a contingent document, then, what kinds of documentary entities—fixed and fluid—constitute the *production* of the Catalogue's data environment? The overall aim of this chapter is to understand how control is initiated and constructed at an entity-level in a taxonomic document in order to facilitate a subsequent discussion about Wilson's concepts of *descriptive* and *exploitative power*.

In the third chapter, "Complex Concepts And Nomenclatural Control," I look at the organization of biodiversity *evidence* and how such documentary evidence is integrally connected to nomenclature, and finally, how nomenclature is controlled through institutional workflows in database environments as a *precursor to taxonomic knowledge production*. I examine the complexity that defines the contingency of (species) *concepts* and the *nomenclature* that biodiversity databases seek to control. As intellectually-circumscribed objects, I describe what types of evidence—type specimens, species circumscriptions, and biodiversity

publications—constitute a species concept, and how the shifting historical application and interpretation of this evidence creates a situation in which multiple, valid species concepts (represented as species names) can potentially apply to the same group of material referents—a situation that creates the existence of, what I call, *complex concepts*. Next, I examine how these *represented documents*, translated into complex concepts, become circulateable, *named* entities in order to maintain intellectual purchase in taxonomic communication. Thus, the management of nomenclature becomes a key focus of my attention (see Figure 5). Nomenclatural control moves us from Patrick Wilson's descriptive power to an ability to *exploit* these names in scientifically functional ways. Managing these complex concepts is far from straightforward, requiring workflows that *both* disambiguate code-compliant name-tokens from non-compliant forms, and that assess the valid name form (and the concomitant species concept attached to that name) among a host of names that potentially refer to the same species concept or type specimen. A current initiative by the Catalogue of Life, known as the Catalogue of Life Plus, is introduced as a possible organizational workflow and administrative process aimed at controlling the iterative disambiguation and articulation of this name space.

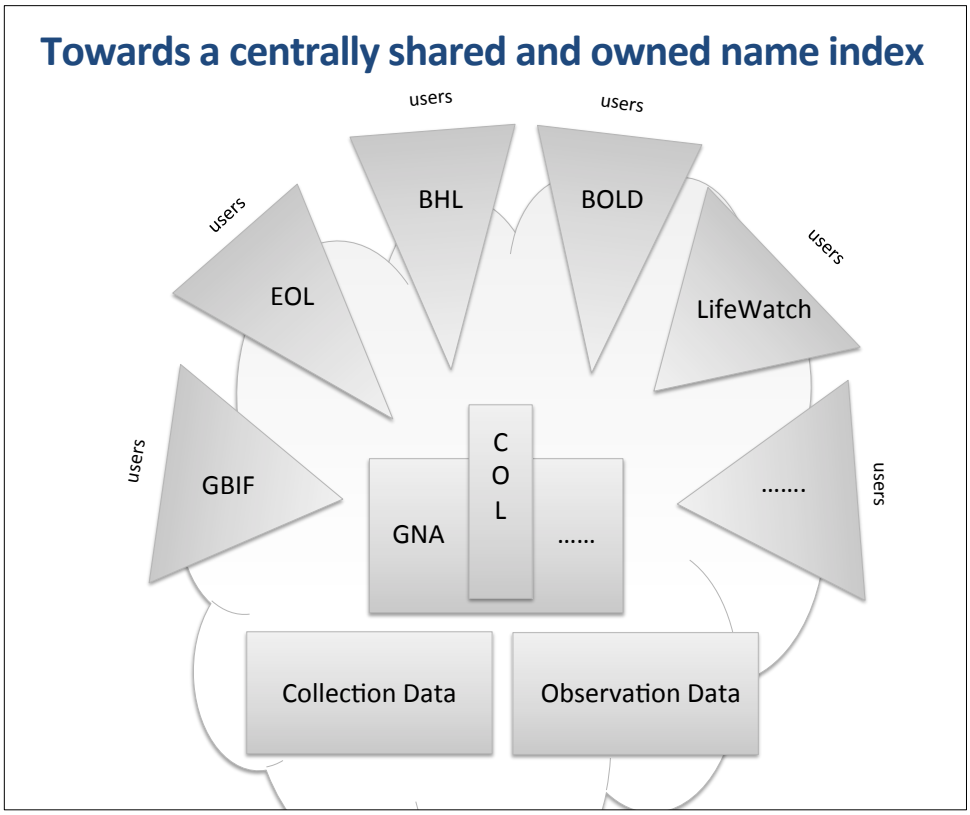


Figure 5. Toward a centrally shared and owned name index. Source: Thomas Orrell and Peter Schalk (2016) This figure illustrates how the Catalogue of Life stands at the center of the iLife environment, providing a stable and curated species checklist and consensus-based management taxonomy that is then utilized in any number of online data collection and access platforms. The Catalogue, thus, stands as the mediator between an undifferentiated pool of names (produced through data collection and observation data activities) and the rest of the iLife environment.

Chapter four, "Documentary Instruments: Taxonomic Specifications, Consensus, and Interpretive Flexibility," then delves into the Catalogue's consensus-based management classification specifically, how it is constructed, and how it differs from the taxonomies that it ingests. Again, taking one of Patrick Wilson's notions—that of bibliographic instruments—I situate taxonomies as particular modes of access that have implicit *documentary arrangements* that define the kinds of work these taxonomies can perform and the ways in which we can understand and navigate the classificatory system as a self-contained ontological and classificatory space. I then introduce what Jonathan Furner (2009) calls *descriptive-oriented* and *retrieval-oriented* biodiversity systems as two key notions to differentiate in this taxonomic

space. The fragmentary biodiversity world is defined by numerous descriptive-oriented taxonomies that each argue a specific classificatory point of view—each arrangement is a hypothesis about how relationships between organisms are defined, related, and subsequently represented. Some examples of specific descriptive-oriented classifications are examined, including evolutionary taxonomy, cladism, and pheneticism. I then look to the Catalogue's management classification as an access-oriented classification system designed, as it is, to reconcile and comingle numerous subsidiary databases for the predominant purpose of access. The chapter ends by extending Wilson's two powers—those of descriptive and exploitative power—to include *extensive power*, which describes classificatory instruments *designed* to be repurposed and reformulated in a number of external environments.

The concluding chapter, "Knowledge Bases, Taxonomic Change, and Contentions with Consensus," looks to the potentialities of the Catalogue as a knowledge base for taxonomic professionals, as well as to its limitations as an extensible taxonomic document. As a taxonomy intended to *function* in particular ways—to aggregate *all* extant data from numerous global sources—the Catalogue's *evolutionary informatic* potential is discussed. The Catalogue as a knowledge base can be used to answer questions about the historical development of taxonomies as a *taxonomic concept repository*. Much like the tracking of nomenclature over time helps us understand the relation of taxonomies to historical biological evidence and species concepts, being able to track the evolution of taxonomic structures can help scientists understand evolutionary trends and make large-scale phylogenetic inferences. I then turn to the qualitative taxonomic transformations that taxonomies undergo as they are brought in to the Catalogue's schematic, and how we can understand such change in light of theories in Information Studies. Such changes will set the stage for a series of critiques about the Catalogue. Joseph Tennis's

notion of *second-order classificatory theory* is extended outward to the Catalogue's documentary space to frame such schematic change, focusing on how classificatory relationships are affected and reconstructed in the process of taxonomic collocation; how taxonomic schemes interoperate within the Catalogue's structure; and how taxonomies are fundamentally transformed as they are absorbed into this global space. Finally, I look to critiques of the Catalogue as a knowledge base and taxonomic model that comes into tension with biodiversity practice, and so look to what its extensive limits are in daily practice. I focus on a series of critiques that include, (a) the unsupportable funding model of the Catalogue; (b) the difficulty with which you can assess the completeness and quality of its component parts; (c) the rate at which data errors are proliferated given the Catalogue's deep integration into the iLife ecology; and finally, (d) its inability to absorb name-tokens that fall outside of Linnaean nomenclatural traditions, such as genetic barcodes.

Conclusion

No biodiversity taxonomic platform can serve all needs for all constituents, the question becomes how global *control* can be balanced with the *flexibility* required to do biodiversity work at the local level. As Broadfield claimed,

> Several suggestions are made as to how we should think of the communal mind whose activity is manifested in the consensus of scientists and philosophers. The relation of a community to its component minds is that "of a whole to its parts," a relation that can be understood by considering the functional relation of the parts of an organism. …But the relation of an organism to its members is not that of a physical whole to its parts, for the latter is a relation in which the parts are not distinguished by differences of function. The parts of a divided whole do not function to make up the whole, whereas organs function in ways which interconnect them in the being of an organism. The relations of these organs are misconceived when we try to think of the organs as parts of the whole, for the latter is a quantitative relation of physical things…The "compositeness" and "aggregation" which are thought to make one mind out of many imply the heaping together of physical things, which does not help explain an intellectual unity" (Broadfield, 1946, p. 76).

The function of this dissertation is to better understand how intellectual unity is problematized in the biodiversity space. What does it *really* mean to say that the Catalogue is a consensus-based

classification? What are its benefits and drawbacks? The whole of the Catalogue is *not* purely the additive qualities of the classifications it includes in its schema. It is both far more and far less than that. Again, the goal here is to think about how the historical, disciplinary, and theoretical specificity of biodiversity infrastructures can inform our own work and theories in Information Studies. The extensive power and consensus-generating capabilities of the Catalogue are especially intriguing in a field where the majority of our documentary systems are tightly controlled and intellectually cohesive spaces—and certainly for good reason, since with that control comes more effective and reliable retrieval capacities. There is still much to theorize about the relationship between classifications as representational entities, and classifications as efficient data communication facilitators; the Catalogue's schematic shines new light onto this longstanding debate. Without looking farther afield than our own discipline, the Catalogue's approach could never force us to question and push against our own approaches to documentation and classification. Examining the Catalogue exposes some of the fundamental *constructed* presuppositions and intellectual impositions we make during the act and practice of classifying. Even the Catalogue's classification of 'real phenomena' shows us that there is no such thing as a *fully accurate* and just representation of the world (for there is no such thing in the classificatory world); there are only *interpreted* documents. Embracing this notion is emancipatory in its own right.

Lastly, this manuscript, I hope, will show us how the theories of Information Studies are applicable in realms far beyond our typical systems of concern. So while it is certainly true that Information Studies is a relatively new discipline that borrows heavily from the philosophical, linguistic, and mathematical domains, among others (Furner, 2004b, 2015), what Information Studies offers other disciplines is the space in which concepts and theories can be *applied* to the

practices of information institutions and information endeavors through various technological (broadly conceived) means. Let this be a proposed model for how Information Studies concepts can be aptly applied to related, but historically and culturally distinct disciplinary traditions. Such an approach, I believe, is essential to the longevity and applicability of, not only Knowledge Organization, but the entire field of Information Studies, which has much more to offer the academy than is otherwise exhibited.

Chapter 2: The Documentation Universe

> *To have bibliographical control over a collection of things is to have a certain sort of power over those things; what things, and what sort of power, it is our business to discover and decide. Let us ask first what are the things over which one might have bibliographical control.*
>
> *—Patrick Wilson*
> *Two Kinds of Power: An Essay on Bibliographic Control (1968, p. 6)*

Introduction

In Patrick Wilson's text, *Two Kinds of Power: An Essay on Bibliographic Control*, he postulates a *bibliographic universe* to identity the kinds of "things" available to be identified, described, and related within bibliographic systems, as well as how these entities are related as part of an ontological system. In Wilson's 'universe,' the primary issues of concern were texts—books, documents, etc.—but he also understood the landscape of texts and objects over which bibliographical systems have *control* exceed far beyond such restrictive conditions. The iLife ecology, as seen in the previous chapter, is a highly integrated network of online biodiversity platforms. These platforms control information. They control data. They control the documentation and evidence of the biodiversity world. But before one can have descriptive control over a system of information, one needs to understand what kinds of *entities* these systems contain, represent, and produce. The articulations of Wilson's bibliographic universe—a product of its time and the conditions of bibliography in the 1960's—in many ways still apply. The Functional Requirements for Bibliographical Records (FRBR) schema, following and incorporating many of the same entity categories as Wilson, shows how his influence radiates deeply in a number of contemporary standards and schema within cataloguing and bibliographic theory (Coyle, 2016, p. 3).[25] Thinking about Wilson's approach to the bibliographical world, I

[25] Though the FRBR Final Report (Standing Committee of the IFLA Section on Cataloguing, 2009) makes no references to the underlying literature that influenced its composition, it's hard to imagine a circumstance where Patrick Wilson is not part of a discussion centered on creating a conceptual model for bibliographic records. Karen Coyle (2016), Allyson Carlyle (2006), Richard Smiraglia (2002, 2003, 2012), among others, articulate Wilson's influence on the articulation of FRBR (most often related to the concept of the "work" entity), though others omit

began to examine about how we might think of the Catalogue's biodiversity database 'universe' in relation to this model, concerned as it is with an ever-expanding set of documents and objects that circulate in very specific digital and professional conditions that affect how they are treated as *units* of scientific work.

In other words, I want to broaden Wilson's notion of the bibliographic universe to think about the *documentation universe* of the biodiversity database environment, and how this notion might expand upon, and perhaps problematize, Wilson's basic schematic.[26] How do Wilson's concepts apply to the Catalogue of Life, and even more importantly, what other theories and literatures do we need to cull from in order to understand the Catalogue in all of its representational and digital complexities?

An essential question to ask when considering composite structures such as the Catalogue (and GBIF, and the like) is how their *wholeness* and documentary boundaries are defined at any particular point, given that one of their defining qualities is that they are ever-evolving, never complete, and always in a state of maintenance, refinement, and data ingest. As Elaine Svenonius

---

Patrick Wilson as a key influence in conceptualizing the FRBR model, focusing instead on Seymour Lubetzky and Charles Cutter's influence on the matter (rightly so, for these individuals, indeed, played a big role in conceptualizing bibliographic entities for the practice of cataloguing) (Denton & Taylor, 2007). It is my general opinion that, despite some scholar's acknowledgement of his influence, Patrick Wilson's (indirect, if not direct) influence in this domain deserves more in-depth attention by scholars.

[26] Richard Smiraglia (2002) also notes the need to expand Patrick Wilson's notion of bibliographic families beyond the bibliographic domain and uses the term "instantiation network" to suit this purpose. He states, "With this term [instantiation network] I retain the concept of connectedness, but move beyond even the textual. (I use "instantiation" instead of the more generic "manifestation," to indicate a sense of temporality; an instantiation is essentially a manifestation at a specific point in time)" (2002, p. 7). I use the term *documentary universe* instead primarily because, in this chapter and those that follow, I am not *only* interested in the documentary instances that the Catalogue of Life produces (that is to say, I am not only interested in the texts, exemplars and objects the work of the Catalogue instantiates—though that is certainly part of my discussion), I am *also* generally interested in the kinds of documents that can potentially *inhabit* documentary databases *at all* whose instantiation is quite separate from their representation and circulation in the Catalogue environment. My documentary universe is *all possible objects* that can inhabit the Catalogue with some kind of evidentiary and token-serving purpose, as well as those entities that the Catalogue produces as part of its role as a documentary entity in-and-of itself. My approach here takes a broader starting point than Smiraglia's discussion, though his distinction is certainly useful to this line of thought. Secondly, while my use of "documentary universe" arises from Wilson's use of bibliographical universe, I want to acknowledge Marcia Bates's use of the term "universe of documentation" in relation to documentary production in the various disciplines (2007).

states in *The Intellectual Foundations of Information Organization* (2009), "Willard Quine characterizes the entities encompassed by scientific theory as consisting of the values of its variables. A bibliographic theory can be similarly characterized, its variable being the entities that populate the bibliographic universe…These are the primary objects, abstract and concrete, admitted into a language of bibliographic description and, as such, the fundamental constructs of bibliographic theory" (2009, p. 31). The informational "units" (or, variables, to Quine, as conveyed by Svenonius) and the boundaries of what "items" reside within, and are produced by, the Catalogue, however, are not so easy to assess and articulate within its fluid ecology. What are the Catalogue's limits as a unit of documentary analysis, and how is it that we can deconstruct these limits to understand how the parts intersect and interact? Secondly, and most crucially to scientific practice, is what are the physical entities these databases represent, as well as the intellectual articulations that define how it is we understand them as *bona fide* objects that occupy a space of coherent and conceptual persistence.

As such, this chapter is broken-up into three broad sections: First, I provide an overview of the literature in Information Studies to illustrate how the discipline defines the most basic entities and concepts in the discipline, such as information, data, and documents—entities that are circulating, to some degree or another, within the Catalogue's space. I start at these basic building blocks for two main reasons: 1) to acquaint readers outside of Information Studies with some of the core definitions and conceptual entities of the field, and 2) to place the subsequent extensive discourse about documents and documentation within a solid disciplinary context. Additionally, while Wilson's universe was unconcerned with "units generally smaller than whole texts and copies of them" (Wilson, 1968, p. 19)—primarily because information extracted from context was less useful and more difficult to appraise—in the biodiversity world, the information

within the Catalogue holds intellectual purchase in the scientific community, acting as fully-fledged documentary units in their own right. Thus, beginning with literature from the philosophy and history of information, I illustrate the progression of *information* from mere normalized patterns of data, to information as meaningful and truthful units of systemically circulateable data. I'll then make the bridge from data to document, invoking Jonathan Furner's (2016a) assessment of the issue, which ultimately defines data as a *kind* of document. With that foundation in place, documentation literature is then reviewed so as to understand how documents act as representations and orientations of *evidence*— and, in the case of the Catalogue, biodiversity evidence in particular. This dissertation uses the term *document* as the fundamental unit within the space of the Catalogue: the Catalogue is a database-document that holds representations of other kinds of documents in the form of a species name (that serve as evidence of biodiversity types and descriptions). I trace how it is that I can make this claim. *Form* is key here: the intended end of this discussion is to show that information within biodiversity databases, in order to function socially and scientifically as valid knowledge forms, must iteratively become formed in ways that conform to biodiversity practice. The names that form the collocating foundation for a taxonomic database must have been negotiated through certain protocols to validly represent species concepts and other disciplinary forms of knowledge—this becomes key in the next chapter and beyond.

This leads into the second movement of this chapter, which describes how *contingency* is an essential concept to acknowledge and understand if we are to conceptualize databases (such as the Catalogue) as documents understood within these sets of literatures. This contingency makes it difficult to define the kinds of entities the Catalogue produces. The Latin adjectival form of contingent, *contingere*, means "to happen," thus databases are always in the process of becoming

something else, always "liable to change," and "dependent for [their] occurrence or character on or upon some prior occurrence[s] or condition[s]" (OED, 2017). Though contingency can also connote a sense of accidentalness or chance, the Catalogue is anything but accidental. Change does not equal chance in this environment, for the Catalogue is closely and concisely controlled. It is this attempt to control contingency (through versioning and editing) that necessarily instantiates new documentary forms (versions) arising from the original database-document. The documentary nature of the Catalogue is *emergent* in that its document-ness is defined by a family of fixed productions that together constitute the intellectual formulation we understand as *The Catalogue*. The database document *can* change and yet is able to retain its identity precisely because its qualities and purpose are held together by defined units of control: instantiated document forms that allow us to assess it in 'concrete' terms. The Catalogue's contingent documentary nature is paradoxically dependent upon its concrete *versions*.

Part two and section three of this chapter begins to specifically identify the basic "entities" of the Catalogue—the intellectual *forms* and fundamental constructs of documentary-database theory—that collectively constitute its documentary families (what Karen Coyle calls the "inhabitants" of its emergent bibliographic universe) (2016, p. 13). And this is where Wilson—as well as other entity schema, including FRBR—is pertinent to invoke. *Works*, *Texts*, *Exemplars*, and *Items* as they relate to the Catalogue become the focus of our attention. If information, data, and documents are the individual building blocks of a database, then how do we understand and define the basic *meaningful* forms these units take within the practices of the Catalogue—meaningful in the sense that they define formulations of data-documents that can be seen as *fixed* or *complete* (even if artificial and contingently so) and useful in scientific practice? In practice, these forms are the species checklists and the management taxonomies the Catalogue

produces.[27] To perform this assessment, close attention is paid to the two specific forms the Catalogue publishes to control the management of biodiversity data: its Annual, "fixed" dataset, and its Dynamic, monthly edition. Each of these documentary streams produces entities that differ wildly in their contingency. While the Catalogue certainly changes (and can never been seen as *truly* complete), it leaves behind trails of its production that together *fix* its documentary space.

This understanding of the component parts of the Catalogue sets the groundwork for the subsequent discussions regarding nomenclature and taxonomic classification—two essential formulations that depend upon truthful information and the stability of documentary entities. In introducing an overview of bibliographic models, Karen Coyle states,

> The challenge for us lies in transforming what we can of our data into inter-related "things" without overindulging that metaphor. There are indeed things of interest to be defined for cultural heritage and creative objects, but our universe of operation lacks the precision of, for example, financial data, where every point of information is precisely known, or the calculation of tensile strength in the engineering task of bridge building. What we describe is not easily subject to quantitative testing, and the difference between success and failure is hard to measure (2016, p. xiv).

This chapter defines the universe of things within and produced by the Catalogue so that we can better understand how they inter-relate in later chapters. This, and subsequent chapters, in no way try to oversimplify or overindulge in the "thingness" and "fixity" of the documentary concepts I introduce and discuss. This narrative embraces documentary contingency, for science is nothing if not the generative accumulation of knowledge based on systematic testing, falsification, and verification. There is no reason to expect the totality of our documents to be any more static than the practices that produced them.

[27] I use the plural form of checklist and taxonomy to emphasize that *many* versions of these entities are created iteratively over time as the Catalogue is refined and edited.

Part I: Tracing Units: Information, Data, to Documents in Database Environments

Information to data.

The first order of business here is to address how the literature within Information Studies defines the fundamental unit of "information" within information environments, and how such concepts might be differentiated from data, documents, and knowledge as they are circulating within the taxonomies and database environments of the Catalogue. As we will see, "information" as a concept is defined in many different ways and takes on many definitions, ranging from a rudimentary, undifferentiated notion of information, to well-formed, meaningful informational units. Using the development of information as a concept as an analytic frame, we can then broaden this discussion to be more specific about how such units are functioning within the Catalogue specifically, as well as within the discipline of taxonomy and biodiversity in general as biodiversity data proceeds to biodiversity *knowledge*. Classification structures are complex in their own right, so as to avoid any ambiguities at that bird's eye level, it makes sense to define what *things* these classification schema are recombining as they negotiate taxonomic arguments. Such a framing is not atypical in information and documentation studies. In the "task of collectivizing knowledge," "order, marking, and selection [are] three essential steps in intellectual occupations," Briet commented in regard to the techniques of modern documentation systems. Such actions, however, as Hope Olsen so carefully indicated in regard to the act of labeling and naming within classification structures (2002), gloss over the complex natures of the phenomenon and entities that we try our best to organize. What are the *kinds* of things that we are ordering and what do they represent intellectually in relation to their scientific context?

As Jonathan Furner indicates, "questions such as ''What is information?'' and ''What is a document?'' received close attention in the 1990s in the field of Information Studies, and the

various suggested answers to these questions continue to be treated as candidate cornerstones of emergent theoretical frameworks in the field" (2004a, p. 234) (See also: (Bates, 1999, 2006, 2009; Buckland, 1991; Capurro & Hjørland, 2003; Floridi, 2010, 2011). These questions are central to examinations of information systems, especially since the goal of any classification system (and certainly within biodiversity taxonomic systems) is to properly locate and describe objects for recall, retrieval and use. Certainly, Information Studies is not alone in using information as a core element of analysis. Furner usefully examines the treatment of the term in disciplines outside of Information Studies, including linguistics, computer science, mathematics, and media studies (2014)—and though these domains will remain tangential to my argument and will not be covered in any depth here, it is worthwhile to indicate that other traditions exist in parallel to this discussion. Readers should be aware that "information," as I use it for the duration of this piece, has a particular literature and pedigree that is disciplinarily specific to I/S. And while the topic of this chapter has as its focus the concept of the document, understanding how this concept overlaps, informs and becomes clouded by the similar and related concepts of data, evidence, and records within the information disciplines can perhaps avoid some confusion. Below I will progressively relate how these concepts can be tied together usefully to situate and set the foundation for our discussion of biodiversity concepts.

In Michael Buckland's often-cited essay, "Information As Thing" (1991), he provides a tripartite structure for the "principle uses" of information that are generally of note in the field of Information Studies. The first, information-as-process, is the usage that describes the active process of "informing" as part of the process of communication (information-as-verb); the second, information-as-knowledge, are those facts, subjects, or events that constitute the "message" (Furner, 2004b, p. 439) of communication; lastly, is Buckland's concept of

information-as-thing, which describes the objects (such as "data and documents") that have the potential to transfer information-as-knowledge, as well as facilitate the events of information-as-process (1991, p. 351). Jonathan Furner expands this material version of information to also "designate symbols, signs, or signals, i.e., noises or marks (or even aromas or flavors) that are interpreted in some way by the hearer or viewer (or smeller or taster)"—according to Furner, "utterance," "data," "signal," and "document" are more or less functionally equivalent" (2004b, p. 439).

These distinctions made by Buckland (and Furner) are crucial when considering biodiversity databases (and, I would argue, databases in general) because information in these spaces can (and will) inhabit the quality of being a process, a kind of knowledge, and a token-bearing thing that moves around in a material space. As will be discussed in great detail from this point forward, at any given moment, a core concept such as a *species*, inhabits each of these information states at different times—potentially simultaneously—depending on what kind of action is being placed upon it, as well as what system-based activities it is engaged with at any given moment. A species (represented semantically by a name) within a database *informs* a receiver about a host of descriptive and identifying information and material evidence that defines its boundaries; the name also represents a series of professional standards and bodies of evidence that render this *knowledge* available to scientists as verified, public information; and finally, the name is also a *thing*—a token—that must be formed in such a way that it can travel in computational and technical systems. The complexity displayed by the different states information can embody has significant implications in terms of how the Catalogue system is designed to control these various aspects. Thus, these categories can help deconstruct and situate the purposes of biodiversity concepts (as information units) within the database environment.

In Figure 6, Jonathan Furner supplies a possible schematic—adapted from Luciano Floridi's map of information concepts (2010)—through which we can identify a hierarchy and progression for types of information.

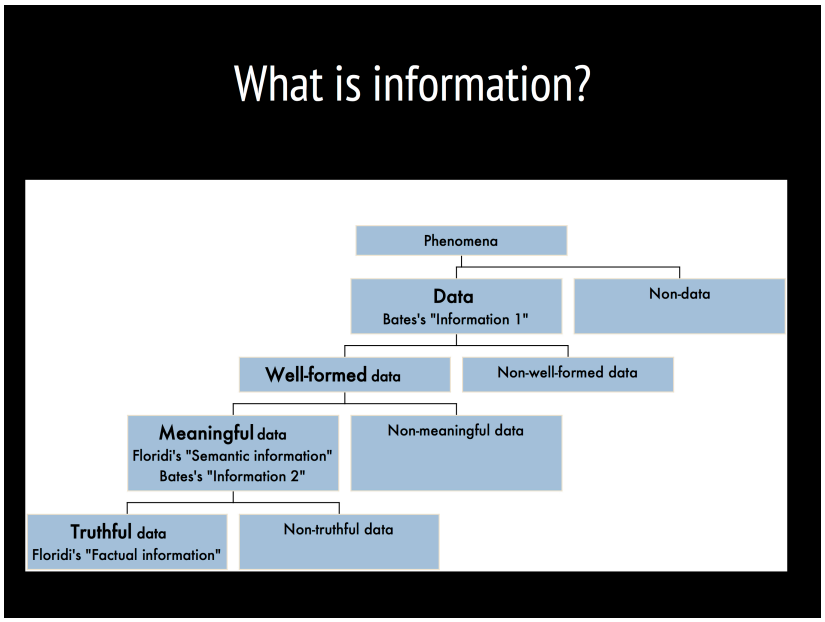


Figure 6. What is Information? Source: Jonathan Furner (2013b)

This graphic model of information will be crucial in our discussion, as it will help us explicate how biodiversity information and data are transformed iteratively through these stages, and most importantly, validated, as it travels through the iLife documentation and information sharing *processes* and *workflows*. This taxonomy of concepts begins at the level of phenomena—processes (natural or otherwise) that occur, can be observed, and can be recognized, recorded, inscribed, or documented in some material form as *evidence* of (or part of) that phenomena.[28] In the biodiversity sciences, the phenomena of interest are ecologies, species, and, most broadly, the natural world. The goal of biodiversity taxonomists is (among other things) to locate, examine,

[28] Of course, the tools, mechanisms, and methods used to record and document phenomena are designed with epistemological stances that influence the contours of whatever data is eventually recorded. In other words, recorded data is never objective data (Gitelman, 2013); it is always *selected* and *manipulated* for inclusion into the system of disciplinary discourse that validate its need for collection. For a nice discussion of how tools, mechanisms and models shape the subjects and objects of discourse see Lorraine J. Daston and Peter Galison's, *Objectivity* (2007).

and define taxa, and delineate their classificatory *position* in relation to a range of other taxon concepts, as they exist undifferentiated in an integrated natural world.

Beginning at the second level of Furner's schematic, then, Data (Information 1) lies at the highest level of extracted information from phenomena. As defined by Marcia Bates, Information 1 is "the pattern of organization of matter and energy" (Bates, 2006, p. 1033)—the ways in which phenomenon create "arrangements that [are] not pure chaos or organization" (2006, pp. 1034–1035). Information here can be seen as a network of connections that, together, constitute the informational area of interest, bounded and extracted from the whole of the phenomena witnessed. Bates's scientific approach to information is particularly inclusive, including "physical, biological, perceptual, and cognitive" (Bates, 2006, p. 1035) patterns, both perceptible and not perceptible, by a human being. Information at this level is merely identified and *potentially* useful, not yet semantically meaningful in that it does not (necessarily) conform to any particular *system* of discourse (and thus is not yet constrained and formulated to enter any structured biodiversity database, let alone a taxonomic schema). One can identify a *potential* species in the natural world, but in order to use this species concept in practice, it must be described, given a name, and connected with other similar formulations, to make sense within scientific and taxonomic discourse. This concept of information roughly corresponds to Luciano Floridi's first component of the General Definition of Information (GDI): that all semantic information is "comprised" of datum from a general pool of pattered information (2011, p. 84).

Well-formed data—the next level down in Furner's schematic—is the first point in which information, as a differentiated object (Floridi, 2004, p. 42), is situated within codes that a) follow the syntactical rules of a system (Floridi, 2011, p. 84), and, most importantly for our purposes, b) are amenable for entry into designed biodiversity schema. Being well-formed

"means that the data are clustered together following the rules that govern the system, code, or language being analyzed" (2011, p. 84). In the Catalogue's realm and the biodiversity world, which deals with species and their organization, this means that our *potential* species has been designated by a *potential* scientific name and that name can now enter the realm of scientific communication (in databases or otherwise). As Bates articulates, if data is not well-formed, then it cannot be *set apart* within the pattern of organization.

Moving downward on the schematic, semantic data (well-formed and meaningful information), according to Floridi, means that "data must comply with the *meanings* of the chosen system, code, or language in question" (my emphasis) (2011, p. 84). In Bates' terms, "Information 2" is strikingly familiar to Floridi's definition, where she indicates that meaningful data is data with "some pattern of organization matter and energy given meaning by a living being" (2006, p. 1036).[29] Without semantic or meaningful data, communication fails (assuming one of the primary goals of information is to inform in some sensible way). In the biodiversity realm, semantically-meaningful data is species nomenclature (and related data) that is validated and proven to follow the rules of codification, *in addition* to following the norms format of a scientific name. Names, and the species concepts they represent, in order to circulate effectively,

[29] A further note on the difference between Floridi and Bates's notion of semantic data is in order here. Floridi's system of information emphasizes a human user as the primary mechanism by which we can understand whether or not "data" is meaningful, truthful, and ultimately, useful. Bates, on the other hand, does not presuppose a human user (recall that Bates's pattern of organization does not, necessarily, have to be perceptible). The Catalogue of Life occupies an interesting space between these two approaches to information: it must balance the social mechanisms set in place to label and order non-human, natural patters of the natural world, but also try its best to maintain a certain adherence to the original order of the taxonomies it ingests to create its management hierarchy. As we will see in chapter four, the Catalogue of Life is a system designed to *facilitate communication and transfer data*, which means it is always *primarily* focused on a potential user as part of its epistemological approach to the kinds of knowledge it purports to construct. In comparison, scientific taxonomies that are focused on a scientific argument are produced to try to argue a certain ontological and classificatory structure about the natural world. Of course, even hypothesis-based taxonomies are intended to *communicate* a certain position to others, but it seems to me, and as I will argue, that the *description* of nature takes precedence in these systems over the *retrieval* of that information for easy use (as in the Catalogue). All of this said, however, all classifications are mediations and intend to communicate information, so all artificial classifications presuppose a user to some extent—even if that user is only the creator of the taxonomy.

must be confirmed as being described according to particular nomenclatural codes and established materially using appropriate type specimens.

Finally, the last level articulated in Furner's schematic is truthful data, which is the final necessary component of semantic information. "Data that are incorrect (somehow vitiated by errors or inconsistencies), imprecise (understanding precision as a measure of the repeatability of the collected data), or inaccurate (accuracy refers to how close the average data value is to the 'truth value) are still data…but if they are not truthful, they can only constitute misinformation" (Floridi, 2011, p. 104). Within the biodiversity realm, only names and identification attributes that are valid and accepted are *true* in the sense that they gain authority and are confirmed (scientifically) to represent a particular species concept, and can be considered known-to-science. Biological specialists and taxonomists spend a great deal of time filtering through scientific names: they must assess the valid name forms within nomenclators, assess their adherence to codes, and compare multiple duplicate name forms to assess which one is the most accurate and correct.[30]

One concept necessary for this discussion that Furner's schematic omits is *knowledge*, which in Floridi's original schematic (2010, p. 20), constitutes the next level downward from factual and "truthful information" ("truthful data" in Furner's terminology). Floridi explains that,

> Knowledge encapsulates truth because it encapsulates semantic information, which, in turn, encapsulates truth. Knowledge and information are members of the conceptual family. What the former enjoys and the latter lacks, over and above their family resemblance, is the web of mutual relations that allow one part of it to account for another. Shatter that, and you are left with a pile of truths or a random list of bits of information that cannot help to make sense of the reality they seek to address" (2010, p. 51).

One point of departure from Floridi, however, is that what constitutes "knowledge" at any given point in the scientific community is subject to consensus-based mechanisms (Broadfield, 1946,

[30] "Truth" is a tricky word to use in this context, for it is always dangerous to claim that any representational structure is truthful. Truthful in this context should be taken to mean that a species concept is "known to science" in the sense that it has been formed in ways that conform to nomenclatural codes. Any proposed, "correct," name for a species concept is tentative and subject to reevaluation.

Chapter 4). Scientific knowledge is *constructed* by scientific communities, it is not self-evident merely because it emerges from a web of what has been deemed "truthful" units of information. For the Catalogue, this is a key component of what makes it a functioning resource for the biodiversity community as a repository of knowledge. The Catalogue conveys informational units that function as both a series of species concepts, traced as they are to their synonyms and variants, *and* a mapping of concepts situated within taxonomic schema that both constitute "integrated" (Bates, 2006, p. 21) knowledge domains that form a conceptual and cohesive unit of delivery. The Catalogue's authority within the scientific community depends on validated, *truth*ful species concepts at its core (in the Floridian sense that they are accepted as facts by the scientific community—that they are species terms that hold purchase). Without this supposition of truth, the reliability of the Catalogue would diminish within the professional community. The process by which nomenclators are compiled and verified (the subject of the next chapter), and then subsequently matched with species concepts, is seen as the creation and articulation of "knowledge" filtered from undifferentiated nomenclatural spaces. But these interpretations change over time, and the Catalogue must account for this change. The Catalogue is about the maintenance and verification of emergent and shifting knowledge interpretations (dynamic and contingent), and as later chapters will discuss, mapping this knowledge over time is no easy task. The Catalogue, then, is not only a "data"-base, but is also a system that must balance various "theoretical constructions" ("schema, n.," 2016), into a fully conceptualized knowledge schema.

Data to document.

Shifting our attention away from information and knowledge for a moment, I now want to chart the relationship between data and documents as articulated in literature. We are, after all, dealing with the Catalogue of Life—a database that organizes various kinds of documents. As

was previously mentioned, documents form the core concept that I will use in this manuscript to describe the fundamental units within the database environment. This is primarily because document theory emphasizes the *evidentiary* value of information, and perhaps more importantly, that a document is, according to Michael Buckland, not "a set of known facts … but that it has propositional content" (Furner, 2016a, p. 300) that attempts to make evident some concept (for example, a species or a set of relationships between species). In order for taxon concepts within biodiversity science and taxonomy to be credible, they must be articulated and circumscribed though a network of evidentiary documents that form the basis of their concept or taxonomic hypothesis. Further, 'document' studies has engaged far more with the museum specimens and biological objects (Otlet & Rayward, 1990, p. 197) that form the foundation of biodiversity work.

Running parallel to Floridi and Bates's theoretical discussions of information and (at least in Floridi's case) data, then, are discussions regarding the distinction between data and documents. David Blair's "The Data-Document Distinction" (1984) is one such text that takes a system-based, information retrieval view of data and documents that is articulated quite apart from the documentalist discourses that find their origin in European writers such as Paul Otlet and Suzanne Briet.[31] Blair's general differentiation between the two entities is that documents, on the whole, are those entities that are most useful to information seekers to make informed decisions (in that documents are what information retrieval systems are intended to return as part of a query), and that data collectively make-up the whole of the document: "Much information germane to decision-making is not data, but contained in documents. Documents often contain

[31] Blair's "The Data-Document Distinction Revisited" (2006) article was released the 2006, the same year Ronald Day published his highly-influential English translation of Suzanne Briet's, *What is Documentation?* (1951/2006), thus the surge of interest in documentation after Day's publication may be one reason this area was not alluded to. Another reason may simply be that Blair comes from an information retrieval background and posits a more system- and technologically-based point of view of documentation. In either case, Blair's discussion is useful to mention.

data, of course, but it is data that has been selected, interpreted and presented for use—it is data in context" (2006, p. 78). Blair (1984) provides four distinctions that illustrate differences in the *qualities* of data and documents insofar as they influence the systemic retrieval of information from search systems; they include,

1. Data retrieval queries are specific, while document queries are "indirect and ambiguous";
2. For data there is a "necessary relation between a formal query and the representation of a satisfactory answer," while for documents, "there is a probabilistic relation between a formal query and the representation";
3. The "criterion for success" is data retrieval is correctness, while for documents it is "utility";
4. Query speed for data depends "on the time of physical access," while for documents it is "dependent on the number of logical decisions the searcher must make in the course of her search" (2006, pp. 79–81).

In 2006, Blair revisits these distinctions and identifies eight more that only serve to blur the difference between the data and document retrieval models. While the differences between data and documents often rest on the specificity, exactness, and scalability of the former (and the ambiguity, indeterminacy, and broad search space of the latter), unlike the conclusion drawn in his 1984 article, in this expanded view, "it becomes apparent that in spite of the numerous differences that we can identify between data and document retrieval, most of these differences stem from a common problem: the indeterminacy of language. Data retrieval and document retrieval do not occupy mutually exclusive classes of information systems, but, instead, are two extremes on a spectrum of representational indeterminacy" (2006, p. 78). For all of the work that Blair invests in *differentiating* "data" from "document" in system spaces, the end result is that, as data and document systems grew in prominence and occurrence, their differences became less significant functionally.

Useful as Blair's examination is, he does little to demarcate a clear distinction between data and documents. Jonathan Furner's (2016a) historical analysis of data articulates how the concept of data relates to the similarly-fundamental concept of document within information studies discourse—and provides a useful, definitive statement about how they are related. In this

piece, Furner concludes that, unlike the prevailing assumption that "all documents are in some sense made of up data" (2016a, p. 288), documents are, in fact, the "primary concept," and "a dataset is made up of documents; and the dataset is a species of document" (2016a, pp. 289, 303). In order to articulate this assessment, Furner exhaustively examines the conceptualization of data by examining its use historically, beginning with classical Latin from the period 100 BCE to 20 CE; tracing its evolution through the more familiar usage of data as evidence (ca. 1648-); to an informational interpretation, which includes the rise of tabularized data that represents "content … about a referent"[32] within the social sciences (ca. 1630-); and, finally, the "computational interpretation" of data as bits (ca. 1980-) (2016a, pp. 290–299). Furner then lays out a logical argument illustrating how one can think about the relationship between text, data, and document, merging "an information interpretation of 'data,' an information interpretation of 'document,' and a subjectivist interpretation of 'information' (as meaning)" (2016a, p. 303). The end result is that the computational approach to data (data as bits) is less preferred than the information interpretation where data, among other issues, is inclusive of multiple formats, substances, and kinds.

Though an obvious statement at this point in database management, within the discourse of biodiversity databases, it should be noted that "data" are nearly universally understood to be of the "informational" type of data, encompassing a variety of different types of objects outside of the numeric form indicative of "computational" data in Furner's schema (2016a, p. 298). As will be expanded upon at a later point, data in the biodiversity world includes two- and three-dimensional images, video, alphanumeric values, type specimens, journal articles, genetic samples, etc. Though, even while data are conceptualized as inclusive of multiple forms in

[32] For a rather exhaustive examination of the use of data in observational scientific activities not otherwise mentioned in Furner's examination, one can look to Loraine Daston and Elizabeth Lunbeck's edited monograph, *Histories of Scientific Observation* (2011).

museum and biodiversity practice, much of the networked infrastructure that supports this work is incapable of more complex data sets, particularly video and three-dimensional data (both due to the memory required to store these instances and third-party software requirements to display these assets). For example, as we saw in chapter one, the media asset management system currently in use at the Natural History Museum, London—Open Text Media Manager (known internally as MAM)—does not have the capability to store video or 3-D files. As Matthew Woodburn, Science Data Architect for the Natural History Museum's Digital Collections Programme, indicated, "we either have to wait for that capability to be put in place or we need to look for a workaround if it's not going to happen. There are [many] kinds of difficulties on the 3-D side of it. Is 3-D an image or is it a dataset? Because effectively it's all bits and bytes but it tends to be ... visualized by a particular piece of software ... We can publish it as a dataset, which is not a problem ... then it would be up to them to find a visualization service" (2016). In more than one conversation with museum specialists, such a distinction was made between 'data that can be used in any particular database' and that data 'that was unable to be properly ingested into the digitization and digital management program.' As such, what constitutes data within an institutional setting at a pragmatic level is often dictated by what formats and types of data any particular system has the ability to accept and integrate.

Document to database-document.

Now that we've made our way from information to documents, we next need to expand on the notion of documentation a bit to understand how they constitute modes of *evidence*—a concept central to the practice of biodiversity taxonomy. I also want to address how it is that the Catalogue, evidence as it is for a variety of organizational, professional, and disciplinary functions, is a type of document in-and-of itself. This latter notion is important, since, as a

documentary entity, the Catalogue produces *other* documents—these other documents, and their relationship to the dynamic Catalogue database, will be the focus of Part Two of this chapter.

In her influential text, *What is Documentation?* (Briet, 1951/2006), Susan Briet begins her exposition asking a fundamental question: how do we define a document? Briet offers the following definition: "any concrete or symbolic indexical sign [*indice*] preserved or recorded toward the end of representing, of reconstituting, or of proving a physical or intellectual phenomenon" (1951, p. 10). This is not a trivial issue when considering biodiversity databases, particularly because the core functions of a database in these terms is to represent intellectual concepts (species/taxon) with some high degree of effectiveness via nomenclatural tokens, in addition to reconstituting (digitally) the intellectual *evidence* that gives these concepts credibility within scientific discourse. In Briet's terms, a document is a *physical* entity (definable, finite), one that, as Ronald Day states, acts as "evidence—of things or larger grouping of things" (1951/2006, p. 48). Paul Otlet also reiterated the physical nature of documents, indicating that numerous objects have the potential to constitute documentary entities, including, "museums and cabinets, collections of models, specimens and samples" (1990, p. 197). Briet's famous antelope exemplifies this fact as well. Materiality plays a key role for documents here, in that these sources of evidence are representational and semiotic *objects* that can subsequently be "placed in a cage," "stuffed and preserved," or have "its voice recorded on disk" (1951, p. 10). The unit of analysis is generally clearly defined in these circumstances: we know what we are talking about when we talk about an antelope—it is a specimen in a zoo, a physical photograph, an article about the antelope.

The material examples used in these instances, however, may give the impression that documents are only *physical* things. Documents are also situated artifacts that are associated

with, and constitutive of, the social and organizational contexts that give them credibility. Briet's antelope, for example, is only an evidentiary specimen because it is displayed in a museum. As Ronald Day points out,

> Otlet's understanding of documentation was expressed through his trope of "the book," which Otlet variously referred to as the book (le livre), the book-document (le livre-document), the document, and generically as "le Biblion." Otlet's trope of the book referred to both the physical object of the book and, even more importantly, to a cultural concept of the book as a unifying form for positive knowledge. Inasmuch as this concept not only embodies the physical object of the book but also is reflective of social and natural "facts," it represented for Otlet a concrete embodiment of the history of true knowledge and is thus a vehicle to global understanding (2008, p. 10).

The use of the word "document" here is a term that is inclusive of, not only physical objects like antelope or books, but also more abstract informational entities that together make a document *informative* as part of a discursive network of other objects, ideas, and norms. This notion of the document as a site for cultural, informational, and material *synthesis*, allowed for a conceptualization of documents far beyond the material bounds of any one object. If documentation could be conceived as a network of associations, then the informational units could be culled from them and rearranged in multiple combinatory ways to formulate new, emergent knowledge sets. As Day states, "Nor is Otlet's notion of the book antiquated by today's hypertext: it was that of a whole with multiple, interconnected parts, a forerunner of hypertextual linking following what Otlet termed the 'monographic principle' (that is, 'atomic' chunks of text)" (2008, p. 11). Paul Otlet and Henri La Fontaine's, Mundaneum—a space in which all the world documentary knowledge could be collected and consulted (Rayward, La Fontaine, & Otlet, 2010; Wright, 2014)—is premised on the very fact that "facts" could be culled from documents, rearranged and reorganized to produce a networked repository of global knowledge.[33] Numerous scholars, including Alex Wright, have rightfully pointed to the fact that the Mundaneum, along with other card catalogue-esque technologies, are the direct ancestors of

[33] Refer also to Paul Otlet and Henri La Fontaine's, International Federation for Information and Documentation, (Rayward, 1997).

database systems (2014, p. 33). Scott Dewey also notes how "Otlet and the [International Institute for Bibliography (IIB)/International Federation for Documentation (FID)] developed precursor techniques and strategies for addressing core problems and challenges of the modern information society in various areas, including systems and organizational arrangements, databases and collections, image databases, database management…" (2014, p. 4). Otlet's conceptualization of documents both as things and facts allowed for the recombination of documentation in databases-like forms.

Following this thread, Matthew Hull notes, "It is unclear if databases, for example, are documents, but they are certainly forms of documentation that demand greater attention in the anthropological investigation of bureaucracy" (M. S. Hull, 2012, p. 261). Geoffrey Bowker, in his text *Memory Practice in the Sciences*, asserts that databases contain the root elements of a narrative that are vital to understanding the development and constitution of social and scientific practices (2008). By producing a narrative of social practices, databases are a form of *documentation*. Much as a paper document—or an antelope—is an artifact representing some *other* state of affairs, so too is a database representative of a social process as it occurred in the past. Richard Smiraglia, in "Further Reflections on the Nature of a 'Work,'" notes, "we speak of a documentary entity…when we wish to speak of a document that has been collected for the purposes of information storage…documents record raw data" (2002). Returning to Buckland's, "What is a Digital Document," he states, "for practical purposes, people develop pragmatic definitions, such as 'anything that can be given a file name and stored on electronic media' or 'a collection of data plus properties of that data that a user chooses to refer to as a logical unit' (1998). Documents *represent* things, just as databases do, but they also act as a "vehicle of meaning: in other words, the object effectively serves some purpose" (Buckland, 1998).

Databases also model, in a functional and pragmatic way, the enumeration and distribution of physical, bona fide, species groups so that certain kinds of *actions* (conservation, policy, taxonomic, etc.) can be employed. Though Buckland indicates that a final definition of a 'digital document' will continue to "remain elusive," for the purposes of this study, given this literature, I will assume a database to be a digital document, given the fact that it is a collection of contextualized data that documents social activities; is locatable in some defined, database (material) form; and represents a fragmented "information object" like those defined by Otlet's (Day, 2008, p. 10).

Continuing with Buckland, "if 'documentation' (a term that included information storage and retrieval systems) is what you do to or with documents, how far can could you push the meaning of 'document' and what were the limits to documentation?" (Buckland, 1997, p. 804). The distinction here between documentation—"tools for intellectual work" in the Briet-ian sense (Briet, 1951/2006, p. 14)—and the document itself—the evidence meant to be organized—merits more examination, especially if we are to understand biodiversity databases as, themselves, a kind of emergent document that, as part of their essential functions, wildly fluctuate in composition as new GSDs are added and redacted; and as changing taxonomic opinions are offered over time. Databases are not merely a mode for "systematic access to written texts" (Buckland, 1997, p. 805); the physical limits and boundaries of database systems are dynamic and permeable in a number of senses. This separation of the machine of documentation (the structural methods by which we organize subsidiary documents; the database), and the collection and organization of evidence (species concepts and *names*) within these systems is a useful distinction, and certainly acknowledged in biodiversity taxonomic circles. Vincent Smith of the Natural History Museum, London, states, "There's the conceptual

aspect to this [taxonomic work] and then the technical implementations, which often try to approximate ideals with practicalities. It's particularly complicated because often we are trying to model concepts, and those concepts are slippery and … difficult to precisely define, and therefore depending on how perfect our information is, it is going to dictate how much [and] what you can do with that data" (2016). One cannot speak of biodiversity taxonomic work without acknowledging the inherent tensions between the document-*system* and the document-*representations* (the data) themselves.

As Buckland has indicated, "definitions based on form, format and medium appear to be less satisfactory than a functional approach, following the path of reasoning underlying the largely forgotten discussions of Otlet's objects and Briet's antelope" (1998). Such difficulties in articulating a clear and concise definition for a digital document arise in the digital and database world, at least in part, from the fact that databases such as the Catalogue of Life are, by design, never complete. Things that "don't have easy boundaries are of indeterminate theoretic status" (2008, p. 144) and difficult to classify, as Geoffrey Bowker has commented. Digital and networked environments force us to re-conceptualize what it means to classify a document-as-entity; they force us to understand that documents are more than just physical formulations, but also encompass a range of community practices, intellectual concepts, and professional guidelines that serve as boundaries in their own right. And most importantly, a key aspect of database documents is that they are often defined by their *instability* and *contingency.* Historically, it is not only the digital environment that prompts us to re-address and broaden the production, function, and definition of "document." Similar theoretical interrogations have occurred within bibliographic and archival spaces as well, domains that have critiqued the traditional object-centric approach to notions of the 'the text' and 'record.'

Contingent Documentary Stability: Bibliography, Relevance, Records

> *All documents are fragments. Let's begin with that premise. Excerpts. All works are partial, almost in inverse proportion to their appearance of completeness … I don't see a simple, positive material fact when I look at a document, I see fields of shifting relations momentarily stabilized in an artifact that exists in a continuum of temporal and spatial and quantum dimensions, only constituted through the framing acts of intervention.*
>
> *—Johanna Drucker*
> *What Is? Nine Epistemological Essays (2013, pp. 48, 58)*

The discussions surrounding the non-linear, indefinite, and combinatory spatial qualities of the database environment have strong analogues in the domain of bibliographic studies, where the boundaries of the documentary and textual entities have, for some time, been questioned. As evidenced by Johanna Drucker's quote above, texts are constantly shifting along temporal and spatial lines. Drucker's (2013) use of Kiernan's study (Kiernan, 1998) of the Boethius manuscript illustrates this concept nicely: wrecked by fire, textual emendations, and additions, the original manuscript text (fragmentary in composition) stands in juxtaposition to the collage-like entity that developed over the years as a transcription was pasted alongside the original. Fundamentally, Boethius' "Consolation of Philosophy by Alfred the Great" manuscript prompts us to ask what constitutes the "text" and "document" in this space of continually emendation and addition? As Drucker tells us, they "are never the same as each other" (2013). The relationship between this circumstance and the ever-evolving database as a textual object are striking. Even as versions of the Catalogue of Life are codified on an annual basis, GSD's are constantly being reevaluated, added, and redacted throughout the years intervening these editions.

This problemetization of the text stands in contrast to traditional bibliographic approaches, such as those espoused by Fredson Bowers, where the notion of the "ideal copy" (an "ideally perfect copy" (1994, p. 113) of a text upon which variations and errors produced by production methods can be identified and documented) is at the core of how we understand descriptive bibliographic practices. As Thomas Tanselle reiterates, "the *ideal copy*" is central to

descriptive bibliography, because it is the element that distinguishes bibliographic description from cataloguing: whereas a catalogue entry, regardless of its level of detail, exists to record a particular copy, a bibliographic description aims to provide a standard against which individual copies can be measured" (1980). Ideality presupposes a printer's intended perfect instance of a book/text upon which all deviations can be measured. The "ideal copy" as a concept, however, falls somewhat short in the realm of biodiversity database-documents. There is no cohesive instantiation (ideational or otherwise) upon which any given version of the Catalogue can be compared against. The "problem of ideality" as articulated by Bowers has certainly been critiqued:

> Even if 'ideal copy' is simply the generic term for a group of objects, there remains an ambiguity: it is either the intellectual ideal (what the printer had in mind) that is imperfectly realized in any particular object (Platonic ideal); or it is merely a nomen, a generic term covering a group of objects: the name can then refer to all products of a press edition and, furthermore, to all as yet undiscovered products of that press-run (Dane, 1995, p. 40)

The same kinds of ambiguities identified by Dane apply in our biodiversity database spaces as well: if there is an intellectual ideal in the Catalogue's case, that ideal is no more specific than "the names of all [*potential*] species set in the context of a taxonomic hierarchy and of their distribution" (Species 2000, 2015b). There is no *ideal* object, only the conceptual aim of *completion*, which cannot be quantified in any stable manner. The standards used to identify and articulate an ideal copy must be an *already-present* set of graphical and/or textual conditions that act as a ground by which we can assess modes of difference. As Jerome McGann noted in *Radiant Textuality: Literature After the World Wide Web* (2001),

> The exigencies of the book form forced editorial scholars to develop fixed points of relation — the "definitive text", "copy text", "ideal text", "Ur text", "standard text", and so forth — in order to conduct a book-bound navigation (by coded forms) through large bodies of documentary materials. Such fixed points no longer have to govern the ordering of the documents. As with the nodes on the Internet, every documentary moment in the hypertext is absolute with respect to the archive as a whole, or with respect to any subarchive that may have been (arbitrarily) defined within the archive. In this sense, computerized environments have established the new "Rationale of HyperText" (2001, pp. 73–74).

Unlike McGann's Rossetti Archive, which he makes a point of saying is "anything but decentered" (2001, p. 74), the Catalogue's database environment is, indeed, decentered and dynamic in *both* presentation (as in *what and how concepts are presented* as the database morphs over time) and in content fluidity (the "text," at any given point, has no ideal cumulative "textspace" (McGann, 2001, p. 149) on which to base any sense of a "center").

Building on this idea of decentered presentations of narrative, numerous alternate forms of sign interpretation can be found in indigenous and New World cultures that collectively exemplify a similar reconceptualization of unbounded document forms. Recent writings on these 'bibliographical alterities,' most prominently by Johanna Drucker (2014), help us articulate how a "non-object" centered interpretation of 'documents' existed far before the advent of digital technologies. Drucker likens the conditionality and "co-dependencies" of electronic documents to the performative and "distributed character of landscape signs, wampum performance, and quipu knowledge" systems as encountered in the sixteenth- and seventeenth-century "contact zones" (2014, p. 21). In these systems, the conceptualization of the document cannot begin by assuming "the existence of a book as an object, *a priori*" (Drucker, 2014, p. 16), but rather the articulation of what constitutes a document is an *interaction* between social conditions, interpretive acts, and whatever form the inscription might take and the communication process might entail. David M. Levy asks, "What are documents?" to which he answers, "they are, quite simply, talking things" (2011, p. 21). But as Drucker makes quite clear, what can 'speak' moves far beyond the "clay, stone, animal skin, plant fiber, and sand" that Levy provides as possible vehicles for communication (2011, p. 23). Meaning, reading, and comprehension within a documentary environment is a constitutive process and performative act between a document and

receiver, occurring within a broader "knowledge ecology" (Drucker, 2014, p. 22) that is defined by a document's use and function at a given point in time.

There is also a sense in which the concept of "relevance" within the sub-discipline of information retrieval is applicable when thinking about document conditionality as well. Documents within a given system will not (necessarily) hold the same significance to any given information seekers at hand. The usefulness of any text, document, fragment, or set of documents, is contingent on a certain set of needs. W.S. Cooper (1971) lays the foundation for this concept drawing on logical relevance as his model, defining relevance as, "whether or not a piece of information is on a subject which has some topical bearing on the information need in question" (1971). Patrick Wilson (Wilson, 1968, Chapter 3, 1973), building on this concept, defined relevance as "a matter of evidential or argumentative status; that which adds to the weight of the evidence for or against an hypothesis is more or less relevant, as it adds more or less weight; that which adds no weight on either side is irrelevant" (Wilson, 1968, p. 44). If, as Briet indicates, documents are entities or objects that provide evidence for the existence of some *thing* or phenomena, relevance helps us understand that "evidence" is contingent upon situationally-specific circumstances and a "particular individual's situation—but to the situation as he sees it, not as others see it or as it 'really' is" (Wilson, 1973, p. 460).

Of course, notable studies have pushed even these claims of logical relevancy and contingency to their limits by embracing a "cognitive paradigm" of documentary retrieval (Allen & Ellis, 2000; Cronin, 2008; Ellis, 1992)—what Wilson had previously defined as "psychological relevance" (Wilson, 1973, p. 458). Such cognitive approaches to information retrieval "sought to develop an understanding of the user's mental models and knowledge states with a view to constructing more effective retrieval systems, stressing agency contexts and tasks

as much as recall-precision ratios and other quantifiable performance measures" (Cronin, 2008, p. 469). Thus, documents are not only contingent in the sense that they can be seen as distributed decentered phenomena, but also in that they can be contingent on the particular (logical and cognitive) need these documents can and are meant to address. Documents (either as singular items or in sets within systems) have no universal boundaries aside from those we fabricate as part of our limited lived experience of them in certain professional, social, and intellectual conditions and contexts.

Distributed records.

Finally, a database can also be conceptualized as a kind of distributed record. As collective entities, archival records that act as "evidence of the functions and responsibilities of their creator" (Society of American Archivists, 2017a), are similar to the way in which databases can be seen as a kind of memory practice. As Geoffrey Bowker (2008) has carefully articulated, databases contain the root elements of a narrative that are vital to understanding the "constitution of social and scientific practices" (2008). Archives and records, like Bowker's notion of databases, play an essential role in maintaining and articulating a "collective memory and human identity" (Schwartz & Cook, 2002, p. 2; Taylor, 1982). By producing a "catalog of traces" (2008, p. 9), databases are a form of *documentation* that "function within shared systems" processually created by their contingent "arrangement with other things … within infrastructural systems" (Gorichanaz & Latham, 2016, p. 1127). Archives concern themselves with "records and documents" that collectively constitute "the archive" (Schwartz & Cook, 2002, p. 5).

The theories behind records and archives have acknowledged the de-centered and distributed nature of electronic records. In contradistinction to a 'paper minded' approach to archives, Terry Cook (1994) articulates the need for new modes of archival collection and

management given the digital "world of virtual, destabilized, [and] fleeting documents" (1994, p. 403). Cook proposes that archivists necessarily need to "shift their emphasis from [focusing on] the physical 'records' to the conceptual 'management' … with redesigned recordkeeping systems" (1994, p. 406).[34] Cook calls for a new paradigm for archival work and theory, one that shifts the emphasis from "the artifact (the actual record) to the creating processes behind it, and thus to the actions, programmes, and functions behind those processes" (p. 410). A key part of Cooks new paradigm—and one that has gained considerable momentum into contemporary archival spaces—is a postcustodial approach to the archival management.

Postcustodial archival theory helps us understand that archivists need not physically acquire and centralize records in order to manage them. The supposition behind this approach is that it "shifts the role of the archivists from a custodian of inactive records in a centralized repository to the role of a manager of records that are distributed in the offices where the records are created and used" (Society of American Archivists, 2017b). The purview and intellectual jurisdiction of the archivist no longer (necessarily) resides in centralized repositories, but instead "sees archivists as regulators, auditors, and 'internal consultants, defining record keeping regimes and tactics'" (Henry, 1998, p. 321). The record—the evidence of, in our case, biodiversity's scientific functions—can now be conceptualized as a disparate, decentered, and networked set of objects that, together, constitute the whole body of documentation under the purview of the archival profession. The South Asian American Digital Archives (SAADA) is an example of an organization that takes a postcustodial approach to archives. Michelle Caswell, co-founder of the organization tells us that, "SAADA is stewarding digital copies of these records for activist organizations, and the communities they mutually serve rather than owning or

[34] See also (Bastian, 2002; Henry, 1998).

controlling them" (Caswell, 2014). The responsibility of the record is coordinated between creator and archivist, *both* within the community as well as within institutional boundaries. Even the record itself can occupy multiple social domains within this model. SAADA's "digitization day" model invites community members to bring their own records and document so that they can be digitized, and with individual permission, those documents are then added to the SAADA repository (SAADA, 2016).

Archives, then, have becomes distributed in their own right—not only in terms of how they understand the purview, authority, and jurisdiction of archival work, but also in terms of how they see the constitution of the archive itself as a set of fluctuating (digital and analog) documentary entities that define an ever-contingent and ever-evolving notion of archival boundaries.

The lessons gained from documentation, bibliographic, and archival literatures, can help us conceptualize and reframe documents, not as static object-centered spaces, but as distributed phenomena that are continually shifting and contingent. Part two of this chapter will now examine how *control* is gained in the Catalogue's contingent space. To do so, we will now dive deeply into Patrick Wilson's entity model and see how it can be expanded to meet the needs of the Catalogue's ever-evolving database environment.

Part Two: Document Forms and Database Entities

Up until this point, this chapter has dealt with the most basic entities within the documentary universe: information, data, and documents. I have also articulated an argument for why databases are kinds of contingent documents, evidence as they are of the formal, intellectual, and social processes that bring them to fruition. But while database documents may be contingent, this certainly does not mean that they lack *any* bounded material instantiations. In

fact, the Catalogue of Life produces a number of distinct and discreet derivative documentary objects. To better understand how we might think about the form of these entities within databases, my analysis will turn to the Catalogue directly to break down what kinds of entities can be located and identified. As has been discussed, Briet and Otlet's approach to documentation places an emphasis on the *function* of the document "rather than traditional physical forms of documents" (Buckland, 1998). The task then becomes how the Catalogue *functions* as a document within the domain of biodiversity scientists and studies. And how does this functionality impose kinds of operational limits and iterative formulations to the composition of the Catalogue-as-document as part of practice? What is actually produced and how can we trace this production along the way?

In November 2016, I was invited to participate in a meeting on the organization of scientific biological names at a global level in Leiden, Netherlands.[35] At this "NAMES in November" meeting (Global Biodiversity and Information Facility, 2016b), the following conversation precipitated between a group of scientists discussing the implications of making nomenclatural data sets available before being edited (and how it may, potentially, contain errors),

> **Participant 1**: Most of the researchers consider this [database] work "in progress." It's never finished. This is part of the problem. It's that the work is not finished.
> **Participant 2**: [I'm] afraid to open raw data to the public because [of our] open data approach. Somebody will take the data without knowing what happened in the work bench. It *is* eventually finished, but it doesn't loop back. [The] Catalogue of Life encourages the publishing of draft systems monthly. [The] final, annual checklist will have a more polished presentation.
> **Participant 3**: [There is] a psychological element. Scientists want it to be perfect but it never will be.
> **Participant 4**: Users don't see how much work [goes] into compiling the database.
> **Participant 1**: Because most of the custodians do it in their spare time. If you have an incomplete dataset you get questions and you have to spend time answering them.
> **Participant 5**: If you care about [the] science you don't want to publish something that isn't refined. [We] strive for perfection because we have the knowledge and want to pass it on. We feel a disservice if it is not finished.

[35] Much more will be said of this NAMES meeting in chapter three where more information about this meeting can be found.

A few important themes emerge from this brief conversation: the first is a matter of the Catalogue's completeness and how that completeness might translate to a matter of quality. It is well understood within the taxonomic community that databases such as the Catalogue of Life are never complete; in fact, in over half of the interviews I conducted and meetings that I attended, this fact was reiterated numerous times. For the Catalogue of Life, the aim is currently 1.8 million species, which means at 1,655,913 species total as of December 2016, they are 92% of the way to their projected goal (Species 2000, 2016e). Scheffers, et. al., estimate that "the completeness of global inventories varies greatly. Completeness ranges from approximately 97% for mammals, 80–90% for flowering plants, 79% for fish, 67% for amphibians, roughly 30% for arthropods and <4% for nematodes" (2012, p. 502). But even if, hypothetically, we were to pin down a definite number of species currently present on the globe, changes will occur as new species are formed, some are driven to extinction, and most importantly, they are reclassified over time given a multitude of variables, methodologies, and emergent scientific discoveries.

The second issue illustrated here precipitates as a solution for the first issue: how to establish and construct (artificially or otherwise) intermittent notions of 'completeness'—monthly and yearly—amid a database environment that is defined by its evolution and instability. To do this, publishing models have been established within the Catalogue to establish authoritative database versions over time. The result of these activities is that many *forms* of the Catalogue are circulating at any given moment: you have "Dynamic" monthly editions of the database that are iteratively posted online, as well as the Annual checklist editions that are published both online and in CD form. How do these documentary forms relate to one another as emergent documents that must be described in information terms (within catalogues and bibliographic/documentary systems)? All issues aside, one thing is clear in this environment: the

Catalogue remains *The Catalogue* regardless of how it fluctuates over time, both in terms of what content is available at any given period, as well as how it might define and express its taxonomic schema spatially and graphically at any given moment. Our task is to understand the subsidiary identities of the entities that describe this cohesive whole.

Within Information Studies, it has always been of utmost importance within bibliographic descriptive circles to be able to accurately represent documentary objects in systems for retrieval. To reiterate, Elaine Svenonius articulates "works, editions, authors and subjects" (2009, p. 31) as the basic entities of the bibliographic universe, and so below we will examine the first two concepts she raises: works and editions. However, the overall model that I will use for this discussion will be borrowed from Patrick Wilson's, *Two Kinds of Power: An Essay on Bibliographic Control* (1968). In illustrating the difference between works, texts, and exemplars, Patrick Wilson (1968) presents the following scenario,

> "A man writes a poem, a letter to a friend, a report on an investigation; he spends a certain amount of time … constructing a particular linguistic object, the piece of language … What he has done can be described in many ways, of which the most important ones are these: has composed or invented a *work*, a poem or letter or report; he has ordered certain words into a certain sequence and so produced a *text*; he has produced marks or inscriptions on some material that constitute an *exemplar* of that text" (1968, p. 6).

Wilson lays out the following entities: *works*, *texts*, and *exemplars* in this example, and later in his text, adds *events* and *objects* to this schematic. I will also borrow from the Entity and Primary Relationship model from the *Functional Requirements for Bibliographic Records* (2009) recommendation to build upon Wilson's final concept, *exemplar*, by expanding this notion to include both *manifestation* and *items*. This seems to me a useful addition to Wilson's schematic given the (relatively) vague presentation of Wilson's *exemplar* concept to mean *both* sets of physical texts as well as individual texts themselves. FRBR's schematic makes a distinction between a *manifestation* as a (potential) group of items, and items as instances that arise from a group of manifested texts. Using these notions, I have summarized how Wilson's concepts

(borrowing from FRBR) can be potentially applied to the Catalogue's entities in Figure 7, below. We will begin with the notion of the work, proceed with the text concept, and then end this section at the exemplar and item level.

But first a quick note: the exercise of identifying entity concepts within the Catalogue database is not merely for theoretical purposes (though certainly that would be enough of a reason to proceed with this thought experiment), there is also a practical reason for this assessment as well: namely that, throughout my fieldwork, it became clear to me through conversations with various stakeholders, that the Catalogue needs a way to understand, keep track of, and articulate the particular mode of data publishing specific to the Catalogue. While they embrace data dynamism on the one hand, fixity is incredibly important as well, especially if the Catalogue is going to be a citable, authoritative resource in the long term. Such fixity also helps apportion credit to the various functions required to build and maintain the Catalogue: schema conceptualization, data wrangling, editorial services, cleanup, etc. The first step to being fully able to describe and provide credit is to understand what functional entities are at work in the database space. Let this be my attempt at beginning to articulate this model.

Work.

It makes sense to begin at the most inclusive entity category in Figure 7, which, in Patrick Wilson's universe, is the concept of the *work*. The work "simply is a group or family of texts, and that for a text to be a text *of* a particular work is the same thing as for it to be a member of a certain family. The production of a work is clearly not the writing down of all of the members of the family, but is rather the starting of a family, composing one or more texts that are the ancestors of later members of the family" (Wilson, 1968, p. 9). If we understand the Catalogue to be one *work* entity, then all of its data derivatives—instantiated in a number of forms—constitute its 'family.' A work acknowledges and embraces a certain degree of change that is inevitable

from successive generations of work instantiations. For a database that is constantly in flux, this is a powerful concept. Recall the limitation of the "ideal copy" or "ideal text," particularly that each of these terms depends on trying to construct a sense of (non-existent, fabricated) stability that is antithetical to the Catalogue's core functioning model (and, as we have seen, such concepts are perhaps antithetical to the ways in which we understand even object-centered approaches to texts). A *complete* text of the Catalogue has never existed since no species catalogue has ever been (and probably never will be) complete. Approaches based on imagined ideal material or narrative conditions for text will not suffice in the biodiversity database realm. A *work*, however, is generative and allows for variations in text over a given space and time.

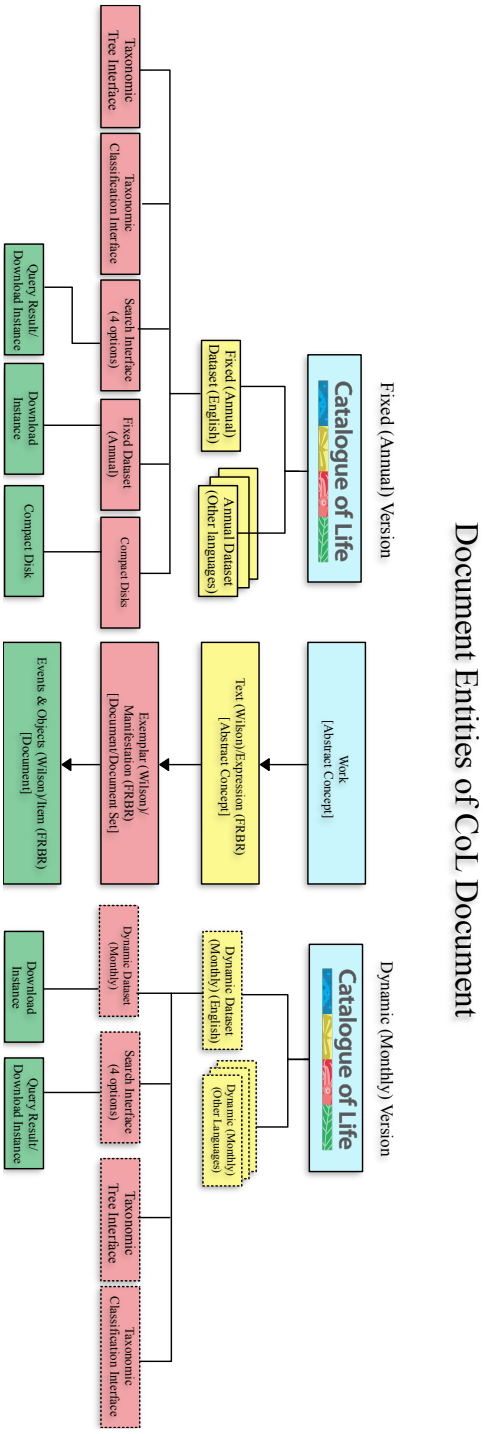


Figure 7. Entities of Catalogue of Life. The center depicts entity types as expressed by Patrick Wilson and FRBR. The Work and Text entities are abstract in that they do not have any tangible or concrete form that can be seen or heard. Exemplars, Events, and Objects are physical and tangible. The left flow chart indicates the entity types for the Fixed (Annual) version of the Catalogue. The right flow chart indicates the entities for the Dynamic (Monthly) version of the Catalogue. The dynamic version is not archived or saved for later use, so they are temporary exemplar documents (indicated by dotted lines).

Discourse surrounding the partitioning of work entities (Svenonius, 2009, Chapter 3) in bibliographic descriptive circles is a more fruitful direction when attempting to conceptualize the Catalogue as a documentary entity. If only because it allows us to understand database-documents as abstracted, intellectual, and administrative objects from which different texts and expressive objects can be produced. Svenonius admits that the "concept of the *work* has never been satisfactorily defined" (2009, p. 35). Lubetzky (1969, p. 33) understood that the work, as a concept, was abstracted and quite separate from the "book" itself. For Svenonius, "the set of all documents sharing essentially the same information" (Svenonius, 2009, p. 35) constitute a "work." And, perhaps this gets us a little bit closer to the way in which the Catalogue can be perceived as a work-like entity.

A fundamental question in these discussions is how one distinguishes one work from another, which is an especially important issue to delineate if our goal is to understand the "limits" (or lack thereof) of a taxonomic databases (Coyle, 2016; Le Boeuf, 2001, 2005). The digital environment presented by the Catalogue presents a number of challenges to traditional notions of bibliographic and documentary entities on top of this already-existing problem. For one,

> The much looser distribution channel of the Internet eliminated the packaging and any vestige of description that those packages contributed. Descriptive rules based on predictable, stable and named "sources of information" (title pages, colophons, etc.) about a resource, with a prescribed order of preference, were not adaptable to resources without title pages or pages, and not suitable for resources that existed in a state of constant change (Coyle & Hillmann, 2007).

Svenonius set-theoretic approach offers the following definition for a work, which is useful: "for book materials (at least), a work W_1 can be defined as the set of all documents that are copies of (equivalent to) a particular document a_w (an individual document chosen as emblematic of the work, normally its first instance) or related to this individual by revision, update, abridgement, enlargement, or translation" (Svenonius, 2009, p. 37). Inklings of the "idea copy/text" lie

dormant in the first part of this definition ("emblematic"), but as has already ascertained, there is no generative Ur text from which successive editions can be usefully compared (in that each edition is drastically different from the former, which, in bibliographic circles, would likely constitute an entirely new work altogether). The second part of this definition, however, is far more useful in this context, for each successive database version of the Catalogue is related by way of revision, update, and enlargement (here enlargement can be viewed in terms of species number and/or the an increase in raw GSD counts). *Relations* are essential here: families of works are held together by *continuities*.

Assessing how the Catalogue work versions relate formally—via commonalities—is certainly a good start toward defining what provides continuity from work to work. On a structural level, what then remains the same? Svenonius's sense that "all documents [share essentially] the same information" is most pertinent here. For one, the *kinds* of information and data objects that inhabit the database space remain the same, which can act as one distinguishing characteristic of a biodiversity database. Standards have been implemented for the Catalogue, which constitute "both the core knowledge set of the Catalogue of Life and around which processes and protocols are designed" (Species 2000, 2014) (see Figure 8). Obligatory standard fields for contributing GSDs include, accepted scientific name, latest taxonomic scrutiny, Catalogue of Life LSID (applied by CoL Taxon Matcher software), and source database. So while the specific species or infraspecific taxa may change (in number, in rank, etc.) among different versions of the database, what does *not* change are the basic metadata fields that structure the information in ways that make it continuous and fully functional within the iLife ecological space.

	Field	Requirement
1	Accepted Scientific Name linked to References	obligatory
2	Synonym(s) linked to Reference(s)	obligatory, where available
3	Common Name(s) linked to Reference(s)	obligatory, where available
4	Classification above genus, to the highest taxon in database	obligatory, where available
5	Distribution	obligatory, where available
6	Life zone	obligatory, where available
7	Current and Past Existence	obligatory, where available
8	Additional data	optional
9	Latest taxonomic scrutiny	obligatory
10	References	obligatory, where available
11	Taxon Globally Unique Identifier	obligatory, where available
12	Name Globally Unique Identifier	obligatory, where available
13	Catalogue of Life LSID	obligatory
14	Source Database	obligatory

Figure 8. Catalogue of Life Standard DatasetField Groups. Species 2000 has defined fourteen field groups to be the standard set of data (version 7, 23rd September 2014) for each species and infraspecific taxon in the Catalogue of Life. (Species 2000, 2016f)

The second significant commonality connecting work iterations of the Catalogue together is what the Resource Description and Access (RDA) (Joint Steering Committee for Development of RDA et al., 2015, sec. 6.2) calls responsibility for the work. The statement of responsibility for a work, "[relates] to the identification and/or function of any persons, families, corporate bodies responsible for the creation of, or contributing to the realization of, the intellectual or artistic content of a resource" (2015, p. 2.4.1.1). According to RDA guidelines, if a change to the responsibility of the work occurs, this is ground for a new description of the title in question (2015, sec.6.2). In his *Principles of* Cataloging, Simon Lubetzky (1969) identified a number of work entity types for which more than one person is responsible for the content; these are:

1. Works compiled by editor from writings or contributions by other authors
2. Works of authorship that have no principle authorship or compiler
3. Works of changing authorship
4. Serials
5. Revisions and adaptations (p. 34-45)

Administered by the Species 2000 Secretariat, the management of revisions, updates, and the enlargement of the Catalogue *work* are controlled by its editorial board. The Catalogue sees itself function under the serials model of production.[36] Executive Editor for the Catalogue, Yury Roskov, based at the Prairie Research Institute at the University of Illinois, Urbana-Champaign stated,

> My contribution to [aggregating names and taxa] was that we need to [approach] the selection and composition of the Catalogue of Life and move it as close as possible to the traditional way of scientific journals. It means that if you have a choice [between] different taxonomic databases that cover the same group, we need to have a peer review process where independent reviewers will tell us which is the best source (2016a).

But a serials approach to defining the Catalogue as a work is just one part of the descriptive equation here, for this model really only applies to that portion of the database that is fixed in iterative volumes (either by Catalogue itself, or by a cataloger who downloads the data set in order to create an exemplar)—namely, the Annual version of the database document. The dynamic, monthly edition is another matter altogether, and far more difficult to qualify and describe. The monthly database set is not permanently stored on any hard drive, nor are the changes made between monthly editions documented in any detail. However, describing these serial iterations take us into the domain of *exemplars* and *manifestation*, which will be described below. First, a brief understanding of the Catalogue's *text* is in order.

[36] The FRBR report certainly acknowledges that, in some cases, "the value for an attribute of a given instance of an entity [in this case the Catalogue document] may change over time (e.g., the "extent of the carrier" for a serial will change as new volumes are issued)" (Standing Committee of the IFLA Section on Cataloguing, 2009). FRBR provides mechanisms by which this change (at least within the expression of the work) can be quantified and described within bibliographic systems, including by describing sequencing pattern, expected frequency, and the expected regularity of the issues. (2009, p. 38). The Library of Congress's "Guidelines for Coding Electronic Resources in Leader/06" (Library of Congress, 2007) also provides a method by which a Textual Continuing Resource can be described.

Text.

Now that we have a sense of what makes the Catalogue function as a continuous work over time, what constitutes the *text* entity in the database space? As Wilson explains, the text is a "sequence of words and auxiliary symbols, is an abstract entity, like the words of which it is composed … The text of which [a] book contains an exemplar is no physical object, has no weight, no physical space" (1968, p. 7). Why this distinction? For one, it allows a text to take on multiple manifestations and instantiations outside of traditional print forms. A text can also be recited, "written, typed, printed, recorded on tape or phonograph record or sound film; a sign-language performance might be filmed" (1968, p. 7); it may also be stored as bits, or "stored in one's memory" (p. 7). Wilson's notion of the text is similar and relatable to FRBR's "expression" entity-type (Carlyle, 2006, p. 265), which is defined as a string of "specific words, sentences, paragraphs, etc. that result from the realization of a work in the form of a text, or the particular sounds, phrasing, etc., resulting from the realization of a musical work" (Standing Committee of the IFLA Section on Cataloguing, 2009, p. 19). The boundaries of the text entity are defined, however, so as to exclude aspects of physical form, such as typeface and page layout, that are not integral to the intellectual or artistic realization of the work as such.

It is the abstract nature of the text of the Catalogue of Life that allows it to inhabit the multiple exemplars that we will describe below. The text of the Catalogue cannot be pre-determined, however, and as it changes over time, the "sequence of words" changes significantly from iteration to iteration.[37] But given the fluctuating nature of the Catalogue, a question to ask is

[37] On average, taking into account the years 2000 to 2015, the Catalogue has changed approximately 16% of its total size annually (Species 2000, 2017d). In the year 2000, the first year the Catalogue fixed an annual version on its dataset, the Integrated Taxonomic Information System (ITIS) comprised approximately 50% of the dataset (Y. Roskov, personal communication, 2017), while in the 2016 annual version (Species 2000, 2016e), it comprised a far less significant 9.7%.

how much change can the Catalogue endure before it qualifies as having a different *textual identity*? Describing Theseus's paradox, Jonathan Furner states,

> The problem of identity over time—diachronic identity—remains a live issue in philosophical debate. The paradox of the ship of Theseus might be familiar in this context. Every day that Theseus's ship is in the harbor, a single plank gets replaced, until after a few years the ship is completely rebuilt: not a single original plank remains. Is it still the ship of Theseus? And suppose, meanwhile, the shipbuilders have been building a new ship out of the replaced planks? Is that the ship of Theseus? (2009, p. 6).[38]

This paradox lends well to the Catalogue's text if one imagines a situation in which taxa are the "planks" of Theseus's ship: as GSDs are replaced over time—for whatever reason, be it because they provide better quality, more coverage, etc.—the source of a taxa might change, but the taxa itself may remain available within the nomenclatural listing and the taxonomic structure provided by the Catalogue. Even more, though the taxa may *appear* to be the same (as in that taxa's *name* is represented in the database) different source databases may, in fact, have defined that species concept differently than another database (different publication sources, term relationships, taxonomic relationships, etc.), and thus, the *knowledge* constituted by these sources has the potential to be appreciably different.[39] For example, imagine the species *Ursus arctos* is contributed by one database, but that database is replaced by another database, also containing *Ursus arcos*. One would need to look to the evidence of that species concept to confirm that they are, in fact, the same species *concept* (not just the same species name). *Text* as strings in the Catalogue are less useful than in typical bibliographic sources in identifying a biodiversity's textual identity—the defining aspects of the Catalogue are much more than its sequence of words.[40]

---

[38] See also Plutarch, 75 A.C.E/2009).

[39] The relationship between species concepts and names is the subject of our next chapter, so more information can be located there.

[40] This is particularly why, in the opening paragraphs of this chapter, I stated that, unlike Wilson's bibliographical universe, in our biodiversity documentary universe we are, indeed, concerned "units generally smaller than whole

While the Catalogue's text may be abstract, it is important to note that the concept of a *work* allows for textual deviations that each constitute one branch of the family of texts. The text is a more narrowly specified entity concept in that it requires a certain degree of (abstract) textual fixity to define its parameters. The boundaries of the text, then, are those that are implemented by the establishment of annual and monthly database sets that have final impressions of textual and numeric strings.

Exemplar and item.

Making our way downward in Figure 7, the *exemplar* and *item* entity levels are where derivative *documents* are produced that can be described, fixed, and retrieved within database and information systems.

Defining the *exemplar*, Wilson illustrates that this concept is what we traditionally understand to be the material production of a text, or as Lubetzky states, the "material embodying the work" (Lubetzky, 1969, p. 55)[41]: a physical book, or a written letter, for example. Elaine Svenonius's term for this entity is the "edition," (2009, pp. 38–43), being the "particular manifestation" of a work. *Manifestation* is the term utilized in the FRBR model to refer to the same general entity type (Standing Committee of the IFLA Section on Cataloguing, 2009, p. 5). A manifestation in FRBR language is the "physical embodiment of an expression of a work" (Standing Committee of the IFLA Section on Cataloguing, 2009, p. 21). Continuing, the FRBR guidelines indicate, "as an entity, manifestations represents all the physical objects that bear the same characteristics, in respect to both intellectual content and physical form" (Plassard, 1998, p.

---

texts and copies of them"—the context of the information contained in biodiversity database has a complexity that must also be accessible and *appraisable* to the system user.

[41] Seymour Lubetzky did not articulate a concept of the *text* in his *Principles of Cataloging. Final Report. Phase I: Descriptive Cataloging (1969)* that was situated between what he called the "work" and the physical embodiments of those works, which he called the *material* (inclusive of "books, film , tape, recording, or other medium containing the work") (1969, p. 33).

20), which means that manifestations can represent sets of objects (though manifestations need not *always* be multiple; some objects, such as unique archival works, have only one manifestation). The Exemplar is *all* available copies or editions of a text that are produced from one instance of Wilson's *text*.

Different graphical arrangements also make for new *exemplars*. After all, one of the qualities of documents is that "the same information" (in this case the *text*) "can be embodied in different carriers" (Svenonius, 2009, p. 113). The FRBR report indicates that, "the boundaries between one manifestation and another are drawn on the basis of both intellectual content and physical form. When the production process involves changes in physical form the resulting product is considered a new manifestation. Changes in physical form include changes affecting display characteristics (e.g., a change in typeface, size of font, page layout, etc.)" (Standing Committee of the IFLA Section on Cataloguing, 2009, p. 22). Further, the RDA framework lists such reorientations of data as an entirely new content type known as "notated movement," which is content expressed through a form of notation for movement intended to be perceived visually" (Joint Steering Committee for Development of RDA et al., 2015, sec. 6.9.1.3). The Catalogue of Life has six browsing interfaces for accessing the data, including: a Browse Taxonomic Tree interface, which allows a user to drill down into the management hierarchy to locate a particular species (see Figure 9); a Browse Taxonomic Classification, which allows for specific searching within taxa levels (See also Figure 9); and four searching functions: searching by all names, scientific name, common name, or by distribution. Each of these new graphical manifestations produces a new data document of the text—a new physical embodiment of the text—that can be downloaded for personal use.

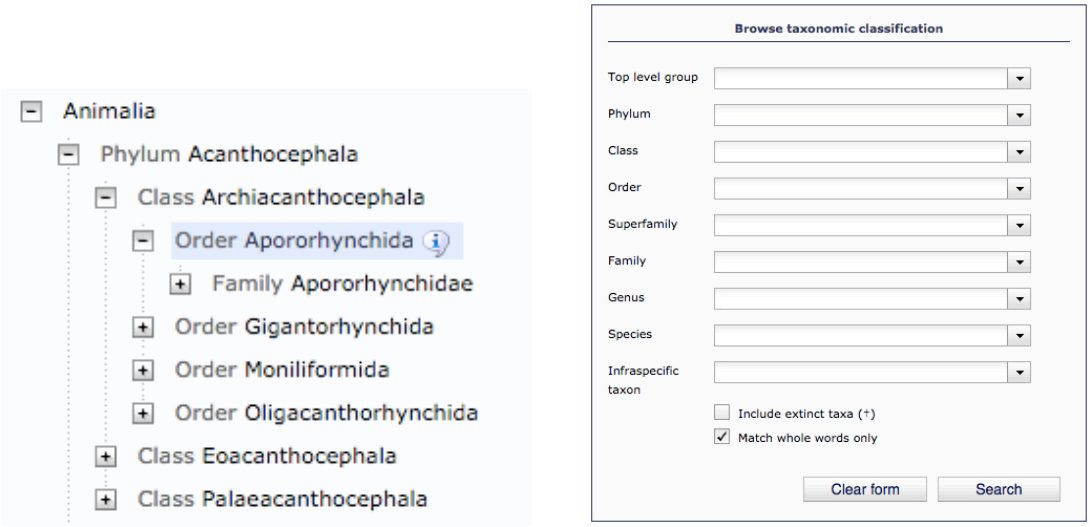


Figure 9. (Left) Catalogue of Life Browseable Tree Interface; (Right) the Browse taxonomic Classification search screen, one of four possible search interfaces available with the Catalogue (Species 2000, 2015c).

The *item*, on the other hand, is the single *object* of the *exemplar* (or *manifestation*). If there is only one exemplar, then it is also the item. Items are generally what we intend to retrieve (or the instances we actually *do* retrieve) when we query some system or seek out a document—some thing (book, recording, dataset etc.) that will satisfy a particular need (unless of course the *need* is to retrieve the entire set of *items* for, say, a comparative study, in which case the *exemplar* set is what would satisfy that need). Often, in the case of the Catalogue, a person is seeking out one instance of a dataset that comprises one document-object—one snapshot of the Catalogue at a given moment in time. For Wilson, *items* can be both events and objects, which means that impermanent performances are distinctly part of Wilson ontology: "if I recite a poem, I utter sounds that are instances of particular types" (1968, p. 7). *Items* need to be *experienced* in some way, and while they certainly need not be physical (like a book), they must be enacted in a certain space within a certain "length of time" (Wilson, 1968, p. 7). This is crucial in a biodiversity database space, for the *item*, within the Catalogue, is only that data that is retrieved within the "performative act" of using the database at any given moment (using any of the browsing mechanisms described above).

On a practical level, the Catalogue has had to manage documentary contingency and constant data change with intermediary moments of fixity. Yury Roskov, Executive Editor, conveyed,

> Another thing which we introduced were fixed and dynamic editions. The Catalogue of Life actually has two products. One product is continuously built and updated … *The* Catalogue of Life, and we release monthly editions. Every month we are changing [the] database [with] some corrections ... from month to month we are growing and changing our content. But again, whether it's convenient for the community of users is the question. So if you talk to young people working with IT technologies they are saying, I need a dynamic resource. If you are talking to scientists who would like to cite resources and cite some particular taxonomic view or taxonomic content or name usage, they need fixed edition. That is why we are making an annual snapshot. So from this dynamic Catalogue of Life, every year we are publishing [a] snapshot as an annual checklist. We started in 2000 and every year we publish this fixed database online (Roskov, 2016a).

The CoL *manufactures* exemplars to make activities like citation possible, as well as to ensure persistent web linkages to the Catalogue in various online infrastructures. Dynamism defines "*The*" Catalogue, but it is the "fixed" versions that provide a mechanism by which we can consider the Catalogue a cohesive, stable, and systematically descriptive 'whole.' The *text* of the Catalogue, in Roskov's statement, is conceptualized as the *contingent* ideal; the *exemplars* make the text stable, functional, and traceable within a distributed digital environment.

Returning to Figure 7, on the left side of the diagram you see the entity map for the Annual, "fixed," version; on the right you see the dynamic "constantly evolving version" (Species 2000, 2015f). The entities at the *exemplar* level and below are the documentary objects of record for the Catalogue of Life, which preserve all of the text associated with those species within the system *at a fixed point in time*, inscribed as they are on different material forms (Blanchette, 2011) and delivered in different graphical data arrangements.

First, the Annual dataset version is available in eight forms: as a downloadable online data set, in a physical compact disk (CD) form, through two taxonomic browsing interfaces, as

well as in four browse-able graphical orientations (described above).[42] In this case, there are only two *exemplars* of the *full* Annual dataset: the CD and the data set available online. The CD version consists of *multiple* object/item entities, which are distributed to individuals around the globe. There is only *one* authoritative instance of the Fixed (Annual) dataset[43] that can then precipitate numerous *items*, or what I have termed 'download instances' in Figure 7. The search interfaces provide an orthogonal entry point into the dataset, where various search criteria can be used to retrieve only those portions of the data required by the user. Each of these queries produces a download instance (*item*). The taxonomic tree interface provides an interactive browseable hierarchy, but no download option is available from this exemplar.

The dynamic, monthly version of the Catalogue, is only available online. But the documentation associated with this version is limited: "anything can change as the [species] list develops: names, their associated details, and their content providers - and there is no tracking of those changes. For that reason, the monthly edition is not the one to quote if you wish to cite a verifiable source" (Species 2000, 2015f). Downloading a full copy of the Monthly edition is somewhat more difficult than the annual version, but possible.[44] Query results from the Catalogue's monthly browsing tools and the taxonomic tree interface are now downloadable are downloadable (as is the case with the Annual version). Two key points about the dynamic version are important to note here: the monthly data set is not archived for future access, nor are the changes made from month-to-month tracked in any detailed way (the ephemerality of these

[42] The CD was initially produced to provide access to the Catalogue to parts of the world and scientific communities that did not have ready and consistent access to the internet, though in reality, most of the users of the CD have been scientists and scholars within the United States (myself included!) who wanted ready access to the Catalogue archives for comparison and citation.

[43] The Catalogue data set is located on the Naturalis Biodiversity Center servers in Leiden, Netherlands.

[44] The Catalogue of Life has to be contacted directly for a monthly version, but this practice is not generally supported since the Annual version is the citable database of record.

dataset are indicated by the dotted boxes surrounding these *exemplars* in Figure 7). Even dynamism is a relative term in this context, for the dynamic version is updated on a monthly basis, and isn't truly 'dynamic' in the formal sense of moment-by-moment updates. This said, it was conveyed to me that (very minor) changes have been implemented in the past through interstitial update periods that were not documented online. While a web page that lists general changes to the monthly editions (back to July 1, 2010) is available (Species 2000, 2016g)—including newly added (or removed) databases as well as database updates—specific changes are not catalogued in any persistent form. Some of these database changes stem from the annual update cycles for particular databases, which are staggered throughout the year. This is why it is especially important that Wilson included both objects *and* events in his bibliographic universe schema. These dataset are transitory documentary *events* of data that require intervention to permanently document. The 'download instance' of the dynamic dataset, of course, can serve as a permanent *item*, but the initial data set as it exists on Species 2000 servers is, indeed, dynamic and unfixed in practice.

Conclusion: Prioritizing Documentary Entities

This chapter was intended to describe the kinds of entities that inhabit the documentary universe of biodiversity databases, and to deconstruct the Catalogue itself as a document to its component parts. Part I provided the terminology for the basic building blocks of the Catalogue's documentary universe, beginning with the concept of information, then moving to data, documents, and, finally, to contingent database-documents. This model will help us situate the increasingly more polished and meaningful nomenclature and species concept informational forms that occupy the database space. Part II then took Patrick Wilson's entity model and postulated how we can understand the *functional* entities within the document of the

Catalogue—identifying those units that help us understand what is unique about its documentary production model and makes it *work* as a circulating resource for scientists. Now with this foundation in mind, the next question to ask is, What is the text of the Catalogue, and how can we better understand how this text is produced? This takes us into the realm of species concepts and nomenclature.

At the terminus of Patrick Wilson's chapter on "The Bibliographical Universe" (1968, Chapter 1), Wilson cautions those that feel the fundamental—and, thus, most important—item within a text is its content, the "separate pieces or items of 'knowledge' that may be found in them" (1968, p. 15). Among the reasons he provides for this position is that some texts are not necessarily valued *only* for the information they contain; some books serve no *factual* informational purpose (a fictional novel, for instance); texts themselves, as cohesive units, present *emergent* knowledge far beyond the sum of the singular pieces of information that constitute their full composition; pieces of information lose some of their value when wrested from their context; and finally, how we understand the *relevance*, *veracity*, and *truthfulness* of a text is entirely dependent on the text's "original habitat" (1968, pp. 15–19).

What we seek, generally, according to Wilson, is not merely information, but the network of knowledge that makes information *inform*ative in its most basic sense. While this is certainly true on the one hand, on the other, *data*bases are textual environments designed for the extraction of selected information that can then be recombined in external environments toward a user's desired outcome. *Sometimes the information and content is all that matters*. Wilson certainly acknowledges this quality of data storage systems:

> It will not be advantageous to make our account of bibliographical control apply generally to units smaller than whole texts and copies of them. This does not mean that we are interested only in whole texts….Nor is it meant to question the obvious utility of assembling all 'known facts' about some range of phenomena into handbooks or compendia of 'data.' It is however, to deny that writings can be adequately viewed as consisting of discrete items of the sort appropriate for handbooks of 'data.' The problem of bibliographical

> control is not simply one of locating items of information, and not one to be solved by attempting to analyze writing of units of information (1968, p. 19).

Unlike the traditional "text" of the documents Wilson has in mind, the text of biodiversity systems are individual *names*—names that arise from a very particular process of nomenclature rules and standardization. In other words, documentary *control* is happening on multiple levels within the documentary universe, not only over the derivative document entities that are produced by the Catalogue. There is a sense in which the units of information within these biodiversity systems *do* have an evidentiary heft quite separate from the lists and taxonomies in which they are embedded. The "statements" (Wilson, 1968, p. 18) the Catalogue contains—species nomenclature and the concepts they represent—do and can subsist without the database text as a whole. But, true to Wilson's assertion, it is not quite that simple. The names that comprise the *text* are situated within a *checklist* that verifies name authority and a *taxonomy* that both facilitates information retrieval *and* situates that name in a larger classification schema (a *kind* of narrative). As a species checklist and taxonomic entity, the text of the Catalogue is valuable precisely because of its completeness as a documentary unit. A certain level of "*relevance*, *veracity*, and *truthfulness"* does indeed come from the context in which names are situated.

In the end, the division between information and text is perhaps an arbitrary one—they are interdependent, comingled as they are in one technically facilitated network. Wresting them apart for the purposes of this examination, however artificial the activity may be, is necessary to uncover the full documentary complexity of the Catalogue's space. The dividing line between the space of the text (nomenclature), and the space of context (classification or taxonomy), is a cause of great schism in the biodiversity world. *How* these two spaces intersect to create a fully-integrated biodiversity system will dictate for whom, and what purposes, a particular system is

useful. The next chapter will describe the relationship between names and the species concepts they represent—the text-space and evidence of the Catalogue. What are names evidence of? How is text accumulated and verified for inclusion into the Catalogue? Nomenclatural "veracity" means something very specific in this context. Furner's information hierarchy (Figure 6) will help us understand this complex territory in disciplinary terms: how undifferentiated names strings become valid forms of scientific discourse. Then, once the role of nomenclature is established, in chapter four, I will delve into the Catalogue's management taxonomy and see how this approach is epistemologically distinct from other, more scientifically articulated structures.

Chapter 3: Complex Concepts and Nomenclatural Control

> *[Suzanne] Briet's notion of documents as evidence can occur in at least two ways. One purpose of information systems is to store and maintain access to whatever evidence has been cited as evidence of some assertion. Another approach is for the person in a position to organize artefacts, samples, specimens, texts, or other objects to consider what it could tell one about the world that produced it, and then, having developed some theory of its significance to place the object in evidence, to offer it as evidence by the way it is arranged, indexed or presented. In this manner information systems can be used not only in finding material that already is in evidence, but also in arranging material so that someone may be able to make use of it as (new) evidence for some purpose.*
>
> *—Michael Buckland*
> *"What is a Digital Document?" (1998)*

> *"Documentary systems" are, at least in some cases, what we would now call in some disciplines "discursive systems." However, the common documentary element of these discourses and their accompanying social networks is that of naming objects according to some institutionally or socially normative systems. In cataloging, objects are placed in relation to other objects based on shared and essential properties and, so, the objects are named accordingly. In formal systems such as library catalogues, indexes, and so on, these names are composed out of formal classes. The relation of the catalogued name to the object is descriptive within classes. In brief, the naming of an object within Briet's notion of indice has a double indexical relationship: the name points to the object and the name reflects the networks in which the object first appears as a named thing, that is, as an example of something, (for example, as a new type within the class "antelope")*
>
> *—Ronald Day*
> *What is Documentation (1951/2006, p. 49)*

Introduction: The Evidence of Organization and The Organization of Evidence

In the previous chapter, the basic units of the documentary universe were described: information, data, documents, works, texts, exemplars, and items. These entities have helped set the foundation for my broader argument about how an information organization system within the discipline of biodiversity studies constitutes what I see as two levels of documentation—or evidence of some "physical or conceptual phenomenon" (Briet, 1951/2006): First, and already discussed, the Catalogue database itself becomes an *emergent* and *contingent document* by virtue of the defining practices, standards, organizational efforts, formal materialities (Kirschenbaum, 2012), and conceptual boundaries that are imposed onto its ontological construction. And how such documentation arises, in part, due to the tension between the contingency of databases and the necessary usefulness and practicality of fixed documentary forms. Secondly, recall Floridi's assertion that information is made up of data (2011, p. 84), as well as Furner's claim that all data

are made up of documents (2016a). One can reasonably extend these notions to state that, information within databases are merely aggregations of documents, and there is no reason to assume that scientific and biodiversity databases are any different. We will now look more closely at the organized documents contained within the Catalogue and how these documents are specifically *represented* and *controlled*.

I will now look inward at the *textual* components of biodiversity systems. That is, what kinds of concepts, intellectual phenomena, and evidence are represented by the text within the boundaries of the Catalogue? I want to think about the evidence *cited* in the Catalogue system, and what kinds of assertions about knowledge are being made by fixing and organizing them within the space of a structured species checklist.[45] Quite simply, we are asking the question, What *are* the documents within the Catalogue, what are they *about*, and are they controlled to circulate effectively? These *represented documents* constitute what I am calling the second level of documentation: those taxon entities that are intended to be described, organized, and accessed within a certain system for retrieval. Fundamentally, in the Catalogue this is a "species" (or more specifically, a species concept), but as I will expand upon below, a species is a document concept far more complex than it initially appears, represented as it is within a computational system referencing numerous external sources (David J. Patterson, Remsen, Marino, Norton, & Page, 2006). A species is both a name string in a technical system, as well as the compilation of evidence that make that name valid as part of scientific discourse and representative of a species concept. The ties that bind these two aspects of represented documents, however, are shifting. *Represented documents* are no less contingent than the *emergent document*. As this work has argued thus far (and will continue to argue more stringently as the manuscript progresses),

[45] The Catalogue's taxonomy forms the focus of the next chapter; here we are more interested in species names and concepts as they are represented by *text*.

emergent and represented documents work together, and recursively affect and influence how the other is defined as each evolves within its own set of practices, standards, and reinterpretations. And they *must* be analyzed separately in order to understand how they function together as an analytic unit, for each concept of the "document" is subsisted by different epistemological and ontological underpinnings and social processes. Each of the interpretive assumptions about these documentary layers operates on a different notion of what the "object" of analysis is and how we can best represent it as cohesive unit.

To this end, this chapter of the dissertation is broken up into two parts: In Part I, I examine the 'unruliness' of the species concept that *names* represent. The phrase *complex concept* is used to describe the difficulty with which names can document the various changes species concepts undergo over a long period of time; concepts, too, are contingent. In informational terms, a species concept must remain a *static* information *thing* in order to *inform* users in a consistent fashion; however, due to the shifting and evolving nature of the species concept as a kind of information-as-scientific-*knowledge*, consistency is difficult to attain over a prolonged period of time. I will illustrate how taxonomists define species concepts (the primary unit of information for biological taxonomies) as an amalgam of three basic concepts: a species name token that *represents* a concept in computational systems[46]; a publication that contains the species *description* as well as other vital diagnostic information to identify the characteristics of a taxa; and finally, a type specimen, which is a physical *type* sample (or set of samples) upon which a taxon description is based that exemplifies a class of real world organisms. These three components of a species concept work together in flexible (yet bounded) ways to create a fluid semiotic system: a type specimen (or specimens) is (are) identified, a type is subsequently

[46] Of course, name tokens are used to represent species in *all* formats; computational systems are emphasized here because of the focus on the Catalogue of Life.

described in a formal publication to delineate a species concept, and then that species concept is codified in a naming act according to strict nomenclatural codes. Problems arise, however, as species concepts change, taxa are reorganized or melded together, as species bifurcate, etc. Thus the limitations of names to track these changes are exposed.

In Part II, I turn my focus to the workflow process by which names get structured into progressively more knowledge-based, authoritative forms based on my field work at the NAMES in November meeting in Leiden—a closed meeting designed to assess the nomenclatural needs in the biodiversity taxonomic sector. If names are complex and contingent concepts, nomenclature specialists need to find ways to organize and track these changes to the best of their ability. I begin this discussion with Patrick Wilson's (1968) notions of *descriptive* and *exploitative power*. I argue that the management of names moves us from the ability to collocate documents (descriptive power) to the ability to use those documents most effectively at the point of usage (exploitative power). I then map-out how the Catalogue validates and codifies nomenclature through a systematic process of disambiguation and concept refinement. Recall Jonathan Furner's schema for information in Figure 6, where he traces the hierarchy of information from phenomena to truthful, refined information forms. Floridi also places an emphasis on the importance of "well-formed" and "semantic/meaningful" data as part of his General Definition of Information (2011, pp. 83–84). Well-formed data is formed in such a way that it corresponds to the particular rules of the system in which it is to be circulated; meaningful data is when a unit of information's structure corresponds to *meaning* (and semantics) within the system. Names follow this same pathway toward becoming *meaningful* information units in scientific discourse; what Thomas Orrell of the Smithsonian NMNH, called being "known to science." Finally, the Catalogue of Life Plus, a newly proposed infrastructure add-on for the

Catalogue is introduced. The Catalogue Plus is designed to manage the complexity of names over time as the pool of potential nomenclature grows, mixed as it is with tokens that do and do not conform to valid scientific name forms and conceptualizations. The Catalogue Plus provides a structured mechanism to *hold* invalidated and alternate name forms, tag them for examination through user feedback, and, once validated, names can be ingested into the core Catalogue for systemic use. The Catalogue Plus is indicative of an exploitative power—the potential for users to contribute to a cache of "writings" that function most effectively at their local level.

Hope Olson once stated, "presumptions have long been made about naming and the language with which we perform the task of naming. Philosophers of language have always recognized the subjective nature of language, but most have identified the resultant diversity to be a stumbling block to mutual comprehension" (2002, p. 5). Scientific names certainly exemplify these issues. Olson is recognized for reminding us that subject representation in library systems matter, and that language is *both* a necessary vehicle for access, as well as a homogenizing agent that obstructs subject complexity in smoothly-functioning systems that construct seamless ontological realities (Olson, 2002, p. 238). Mutual comprehension and communication of biodiversity information is certainly the goal for systems like the Catalogue, and while strict rules have been articulated to *control* the name-space as objectively as science can articulate, the concepts they represent are subjective by their very nature. This fact is not lost to any practicing scientist intimately familiar with the standards and processes of their professional craft; it is part-and-parcel of *doing scientific work*. Olson proposes eccentric techniques of classification in *The Power to Name* (2002, pp. 238–240), which include permeability and dynamism. The Catalogue, and the nomenclature environment that now surrounds it, then, exemplifies one way in which the scientific community is articulating a

recursive technique aimed at solve these long-standing, acknowledged limitations on scientific nomenclature.

Part I: Documents Within Documents: Unruly and Complex Concepts

In September of 2016, a call for papers was delivered to my email associated with a working group session titled, "Resembling Science: The Unruly Object across the Disciplines." In the description of this meeting it was said that,

> We might even argue that the history of consolidating and communicating scientific thought is structured by a tension between two kinds of unruly objects: the objects we seek to represent, and the objects produced by representational media. Scientific media instantiate a wide range of representational modes, from drawings, tables, and diagrams to printed text and script in various languages (Latham, 2016).

As Geoffrey Bowker (2008, Chapter 4) makes readily apparent, the systems that we create to represent and circulate the collective memory of biodiversity science are unruly at best—fragmentary, incomplete, and forgetful at worse. The mechanisms that biodiversity professionals craft and implement to circulate knowledge in space and time—in Bowker's case, he focuses on scientific names and classification—are difficult enough to manage within the domain of biodiversity itself, let alone as these mechanisms get absorbed into other domains of discourse that further obfuscate and rearticulate how that knowledge is constituted in the first place. Biodiversity database nomenclature and taxonomies are such important spaces of discourse because they strive to mediate, locate, and fix unruly species concepts *within systems* that are fundamentally also plastic and sites of intellectual contention. The unruly-ness of scientific infrastructure arises, in part, because science itself is a process (D. L. Hull, 1988) that depends upon the progressive building of knowledge over time (D. L. Hull, 2001, p. 225).

The unruly nature of concepts within the database realm became readily apparent when, in April of 2016—at the very beginning of my fieldwork—I was invited by the Catalogue of Life Global Team to present my dissertation work at their Annual Symposium in Crete, Greece. One

of the central themes of my presentation was how we could begin to think about what constituted the "work" of the Catalogue of Life—what its boundaries were and how we, in Information Studies, can begin to think about its identity as a persistent, published object on the one hand, as well as a dynamic resource on the other. Naively, a comment I made focused on the volume of data that was being aggregated in these systems, and how that volume was one impediment toward *control*. At the end of the presentation, a prominent taxonomist raised his hand and commented, "The distinction here is that we don't just deal with just big data. We deal with *complex* data, *complex concepts* that change over time. The question is how you capture that process" (my emphasis). What this taxonomist was hinting at was that my model of "the work" was predicated on the notion that the *work* documented concept-objects that could be easily located, and that the data elements were relatively stable throughout the process of organization and information collection. It had nothing to do with the quantity of data—or least it didn't insofar as even smaller sets of data present the same problems of complexity and change—but had more to do with the balance between providing systemic flexibility (to account for the shifting nature of concepts) and the fixity required of a system used for organizational purposes.

This response got me thinking about how taxonomists define complexity as a part of their practice. How could I locate and unpack that complexity within the biodiversity management process? Particularly as it relates to the fundamental unit of information in biodiversity work—the species—and how the process of its articulation interfered with the basic, procedural aspects of data management. What is complex about the process of articulating biodiversity concepts that makes them so difficult to manage in a technical environment? What follows is a breakdown of what constitutes the species concept and how it becomes adhered to a certain set of documentation. The result is such that the document of the database (a name string) is comprised

of external documents upon which its validity and constitution depends, but these external documents do not adhere to a stable network of related meaning.

Material stabilization: Types.

On a very basic level, biodiversity databases are trying to locate, track, unify, and represent one of the most fundamental discourse units, as well as one of the "discipline's oldest and most vexing [intellectual] problems" and constructs (De Queiroz, 2005, p. 196): the species. In practice, identifying this basic unit of analysis is no easy task for its concept is subject to expert construction. 'Species' in practice are an amalgam of basically three different intellectual formulations which we will discuss sequentially (the 'botanical trinity' as one scientist called it)[47]: a *type* specimen, an individual instantiating a *class* of entities (Daston, 2004), that acts as a material ground and "the name-bearer of the specimen associated with a name by the act of description and publication"); a publication (that includes the circumscription, or description, of a species); and, a name (articulated based on a series of rules codified by international commissions) (Winston, 1999, p. 173). "Specimens are the basic operational taxonomic unit (OTU)" of (Berendsohn, 1995, p. 208) systematics and biological classification. The term "operational," arising from the tradition of numerical taxonomy (Sneath & Sokal, 1973), is used in this context because species represent the "evidence and operations used to recognize species in taxonomic practice," (K. D. Queiroz, 1998, p. 59). [48] Species as operational units are "*testable* concepts" (Wiley & Lieberman, Bruce S., 2011, p. 29, my emphasis), concepts that are

[47] While there are a number of differences between the botanical and zoological codes, the "trinity" still stands for zoological species concepts.

[48] In the context of the many meetings and discussions I've had in fieldwork, OTU as a representative token for species is used *both* specifically (to mean an operationalizable concept arising from numerical taxonomy and phylogenetics), as well as *generally*, to mean a unit that is based on a set of hypothesized circumstances (see discussion of Catalogue of Life Plus, below, for extrapolation of this general use of the term).

contingent on the repeatable articulation of a species at any given point and time as sources of evidence and interpretation change.

On the one hand, it is easy to think that the primary *document* of biological taxonomy is the *type*—the *one* physical thing that presents itself as a stabilizing node that acts as evidence of the existence of some external group of biological individuals. A type is Briet's ubiquitous museum-residing antelope, representing the *concept* of antelope in the wild that can then be circulated in any number of environments: museums, databases, zoos, etc. After all, when a "species" must be verified in practice, one must "look at the actual specimens of the potential members of the species in question" to truly remove any doubt, because as Smithsonian Research Associate, Judith E. Winston states, "There is no substitute for the specimen" (1999, p. 96). Types are the source material that "settle disputes over species identification" (H. C. J. Godfray, 2007), and also the voucher specimens upon which molecular examination of species are based (Seberg et al., 2016; Smithsonian Institution, 2017).

Type specimens also regulate taxonomic nomenclature (Witteveen, 2015, p. 570) and prevent the chaotic inflation of names that occurs when there is no stabilizing mechanism for their application. The type is nowhere near "typical" in the sense that it represents the variations inherent in an external organismic class (Daston, 2004; Ereshefsky, 2007, p. 261; 1988, pp. 496–498). When Linnaeus began using binomial nomenclature, the formal identification of types was not standard practice, though Linnaeus certainly based his descriptions on samples (Ereshefsky, 2007). It was not until the late 19th century that types entered into common usage, along with the codification of types in various codes of nomenclature (Daston, 2004, pp. 159–161). Daston identifies the early twentieth century as the period in which "types" start getting cited in publications and journals (2004, p. 160) to ground concepts descriptions. Initially, the

assumption that the type *be* typical led to an inflation of the application of types, as well the frequent replacement of types for newer exemplars that more adequately matched phenotypic observations (D. L. Hull, 1988, p. 498). "The resulting confusion led systematists to rule that once a specimen is designated as the type specimen of a species, it can be replaced only in cases of duplication or accidental destruction. The sole function of the type specimen became to serve as the name bearer of its species" (1988, p. 498). The unbreakable bond between type and name is clearly articulated in the *International Code of Nomenclature for algae, fungi, and plants*, by the formal phrase "nomenclatural type" to describe the type specimen: "the application of names of taxa of the rank of family or below is determined by means of nomenclatural types (types of names of taxa)" (2011, sec. 2.7). The result is that regardless of how a taxon is described over time, it remains what Joeri Witteveen calls a "necessary truth that the taxon's type specimen falls within its boundaries" (2015, p. 569).

Type specimen collections are "some of the most precious holdings of major natural history museums around the world," but relatively little seen in public arenas (Daston, 2004, p. 158). Types form the core collection of institutions and are kept indefinitely so as to maintain control over the nomenclatural space (see Figure 10). Specimens also stand at the center of sound and verified taxonomic work, playing a major role in *correctly* (as in, according to stated rules) assigning names to taxa, or reassigning a species to an updated location in the hierarchy. As Timothy Utteridge, Head of Identification and Naming and Senior Research Leader at Royal Botanic Gardens, Kew, articulated,

> [Taxonomy] is not about just names. We use names to identify *real things* in forests. So behind every name there is an organism and vice versa, so to know what that organism is … you have to go back to the original [type and descriptions]. The original sources. And what is great about [the Biodiversity Heritage Library], [is that it's a] camera copy ... it's *there* you can see it ... It feels [like] something that is not being fiddled with. When you look at BHL you are fine with that; but when you look at a plant list you might think that they scan the page number in wrong. So, even though the plant list might have the full citation, you will still click that link to take it to BHL (2016).

As is elicited in the narrative that Dr. Utteridge presents, aside from its purpose in controlling the production of names, a type also bridges the abstraction of concepts with the material and natural concerns of taxonomy—the species themselves. In another, unrelated conversation, a scientist remarked, "types remind me what I do this [biodiversity] work *for*." Secondly, Utteridge also implies a sense in which the physical material of species is more verifiable and trustworthy as a kind of circulating knowledge within the discipline, rather than that of biodiversity database. Databases, effective though they may be in storing and communicating information, provide no mechanism to directly assure the accuracy of the information they contain. Mistakes occur within database environments, and secondary compilation sources can never serve as a replacement for the original types and supporting documentation.[49] Taxonomy in the professional scientific world is a process of manual verification, and in the database environment, there is no way to immediately judge the veracity of information with such distance between the *documented* instance of evidence and the evidence itself. Certainly, some mechanisms have been implemented to assist in this verification process, such as indicating the date of the "latest taxonomic scrutiny" (Species 2000, 2017a) or providing a quality indicator status for taxa information. This is one reason why it is so essential to have all database content linked back to original source; by doing so, it provides the bread trails necessary to locate and verify contributed information. Type specimens provide the most authoritative means available to ensure the production of valid and comprehensive taxonomic knowledge at a local level.

[49] This potential unreliability of taxonomic and nomenclatural databases is explored in much more depth in chapter five.

Figure 10. (Left) Original East India Company Type Specimen Cabinets. (Right) A type specimen folder from the East India Company Cabinet. Royal Botanic Gardens, Kew. Photo by author.

Documentary warrant: Publications.

The second component of the species concept 'trinity' is the publication. While type specimens bear the name and control the inflation of its use, the publication provides many qualitative data points of interest, including the description of the species (its circumscription); known synonymic variants; type material examined to ground the description; a diagnosis (the identification of characters that differentiate a species from its relatives); if necessary, distribution information and illustrations, as well as a host of other potential descriptive information points (Winston, 1999, pt. 3). Very particular requirements must be met in order to qualify as a valid publication that places libraries and other public documentation institutions at the center of biodiversity and taxonomic practice. Within the ICZN, in order to qualify as valid, publications must "be issued for the purpose of providing a public and permanent scientific record" and "be obtainable, when first issued, free of charge or by purchase" (1999, sec. 8.1). An amendment was later added, due to the introduction and proliferation of digital media, that a

publication must also encompass "widely accessible electronic copies with fixed content and layout" (1999, sec. 8.1.3). In order for names to enter the communication stream of scientific discourse, they need to be within the domain of "public knowledge" (Wilson, 1977) so that any and all elements related to a species concept could be confirmed. Good biodiversity and taxonomic work rests on Patrick Wilson's notion of the "complete library"—a biodiversity descriptive library "containing a copy of every published record" (1977, p. 87), not unlike the aims set forth by the Biodiversity Heritage Library in their coordination and digitization of biodiversity literature for global access (refer to Timothy Utteridge's comment above about the importance of primary source literature in verifying species concepts). With all the documents at their disposal, taxonomists can then prioritize name instances *and* retrieve those instances for use in taxonomic databases (more on the latter below).

The date of a particular publication is crucial here. Publications also provide a mechanism by which *accepted* names are officially brought into scientific discourse as representing a taxon (D. P. Remsen, 2010, p. 152), and the *imprinted* date on the document is used to assess the priority of any given name over another. "The problem is that, in a widely distributed literature—across all the continents and a number of disciplines, each with their own sets of journals—you need a rule to decide which name you are going to standardize" (Bowker, 2008, p. 159). Even if distributed literature is in different languages, the use of Latinate forms for names helps collocate that information nonetheless. In contemporary terms, finding specific date information is relatively simple and usually quite straightforward, given that bibliographic information is readily available. However, in conversation with zoologists at the Smithsonian National Museum of Natural History (NMNH), finding an accurate date for older publications is often quite difficult. Many of the zoological journals held in the NMNH dating from the early

twentieth-century, for example, have an accession date stamp (the date a journal was added to the NMNH collection), but do not have an imprinted published date on the issue themselves, be it by design or by a missing copyright page. Sleuthing, then, becomes a paramount task for nomenclaturists attempting to assess nomenclatural priority. Interpolating dates from the historical record surrounding the establishment of a date is not uncommon, especially before the implementation of codes that required publications for the instantiation of the new name. Timothy Utteridge describes one such example,

> The genus I work on, Maesa that was recognized by, I think a Danish guy ... [on] an expedition into Saudi Arabia. And then at the same time, the Forsters went around, I think with Cook, and came back with another [sample] and they called it Baeobotrys. And they published this and it's exactly the same time as Maesa. But the only way they can work out which one takes priority is that someone worked out when their ship landed in Portsmouth, how long it would take them to dock, how long the post carriage took from Portsmouth to London; [and] how long the editor would have taken to write it up. So, they've [assessed] to the day, which one has priority and it came out that one [Maesa]. So, sometimes a bit of detective work is required (2016).

In terms familiar to the field of Information Studies, these publications provide the documentary (bibliographic) *warrant* necessary to include any given name as part of the constellation of other accepted terms. The function of such standardization systems, as Claire Beghtol makes clear, is that it provides the classificationist the authority "to justify and subsequently to verify decisions about what classes/concepts to include in the system" so the system "will be helpful and meaningful to classifiers and ultimately the users of the documents" (1986, p. 110). One of the warrant "perspectives" that Beghtol lists is literary warrant, which generally speaking, describes the inclusion of concepts and the building of "classification based on the collection, … the 'literature in the sense of a body of works" (Kwasnik, 2010, p. 107). Warrant, in essence, assures that the units of information included in the system (species concepts on the abstract level, names on the practical, token-specific level) maintain a certain degree of purchase and semantic usefulness within the system. As was evidenced in Utteridge's quote above, biodiversity knowledge databases—and knowledge sources in general—gain their

cognitive authority (Wilson, 1983) from these established standards and their implementation within systems such as the Catalogue.

Name tokens and species concepts.

> *"Naming information is the term I use for creating document surrogates…I choose the word 'naming' because it connotes the power of controlling subject representation and, therefore, access…Theories, models, and descriptions are elaborated names. In these acts of naming, the scientist simultaneously constructs and contains nature"*
>
> *—Hope A. Olsen*
> *The Power to Name: Locating the Limits of Subject Representation in Libraries (2002)*

But the 'type specimen' of the physical world, on which the stability of the *communication* of taxonomy depends, as well as the publication that codifies a species concept's described identity (its circumscription), is functional only in the form of a *name* in database systems (and, really, in a system of any format—print included). Our discussion of both types and publications could not have transpired without the concept of a name to represent the comingling of these documentary units; names bring the data together. As Lorraine Daston states, "Indeed, it is the name of the species, rather than the species itself, that is directly attached to the type specimen" (Daston, p. 162). Names are the vehicle by which "species" become, as Ronald Day states in *Indexing it All*, "meaningful things" within a larger network of relationships *(2014, p. 6)*. "Taxonomists use 'names' as tokens for concepts of species (and other taxa)"(D. J. Patterson, Cooper, Kirk, Pyle, & Remsen, 2010, p. 3), and "are a part of a 'taxon concept' (Kennedy, Kukla, & Paterson, 2005, p. 82). "Tokens are said to instantiate types; they exemplify embody manifest, fall under, belong to types they're occurrences, instances, members of types" (Furner, 2016b, p. 120). Names as tokens, therefore, merit more examination in these taxonomic spaces, given their pivotal role in documenting *species concepts* within information systems. The general assumption might be that, as tokens, names are stable entities, and that each token, being

unique, correlates to a clear set of documents and concepts, but this could not be farther from the truth.

One of the fundamental problems with using names to convey information is that they are incredibly slippery notions, not only in computational environments, but in general: names cannot always easily be paired with the species concept that it is intended to represent, nor can they "be used to unambiguously identify a concept" (Kennedy et al., 2005, p. 82). Even with the advent of globally unique identifiers (GUID) for name strings, it is often the case that specimens and names are given new GUIDs once they are brought into various local systems. Doug Yanega, of the Department of Entomology at the University of California, Riverside, relates such as case,

> Every [Entomology Research Museum] specimen has a global unique identifier, which is on the label, which is a combination of our institutional code plus an actual specimen number. We accommodate anybody's [GUID]. There are databases with different in house things that won't accommodate foreign GUIDs. Then they have to reassign their own internal and then you have two GUIDs for exactly the same thing. If it's globally unique then you don't have two of them! (2016).

Indeed, other taxonomists and biodiversity informaticians I spoke with recognized this general problem as well, and that given this practice, names, by default, become the one binding agent for disparate information. Yet, as Nico Franz conveyed, "names are not good enough" (2016). With the shifting concepts these names represent, even GUID's cannot use machine algorithms to correctly map the connection between species concepts and name transformation over time.

How is this the case that names lose stability over time? As David Remsen, biodiversity informatician and Director of Marine Research Services at the Marine Biological Laboratory in Woods Hole, Massachusetts, notes, "there is no direct relationship … between symbols (i.e., names) and the real-world objects (the referent) they represent. Meaning, or the relationship between the name and the object, is conveyed only through a concept that exists in the mind of the user of the name…. In biological taxonomy, a species name refers to a concept anchored by a

specimen but created in the mind of a biologist … The function of the name is to facilitate communication. Communication is facilitated, however, only when the concepts (not the objects) are approximately congruent" (2016, pp. 210–211). Such congruence is difficult enough to verify within museum repositories where the type is right in from of you, but once you enter names into an *interface system*, such correspondence becomes nearly impossible to verify. Names are neither stable, nor "unique identifiers for taxa" (2016, p. 210).

Despite rules dictating the application of names to taxa, the species concept that it represents is nothing if not static. Names change readily and *normally* as part of scientific descriptive and taxonomic work. Any number of variations and updates can occur to the concept of the species over time. Richard Pyle, Associate Zoologist and Database Coordinator, at the Hawaii Biological Survey and Bishop Museum, jokingly remarked during the 2008 annual Biodiversity Information Standards/Taxonomic Databases Working Group (TDWG) meeting, "taxonomy is the perpetual classification of mis-named species," and noted that this circumstance is "a necessary evil … that is fundamentally necessary for, not only biology, but particularly for biodiversity informatics" (2008). Multiple names can represent the same taxon concept (synonyms); one name can be used for many entirely different concepts (homonyms); and one name can refer to two or more concepts whose circumscriptions overlap (usually resulting when taxa are split or merged over time) (D. Remsen, 2016). And to add on top of this, each circumscription applied to a name and type is, by definition, an approximation (and *operationalization*) "equivalent to generating a new hypothesis in other branches of biology," and such a hypothesis is always open to new interpretations as new forms of evidence or new modes of data analysis (computational or otherwise) are introduced (Gaston & Mound, 1993, p. 139).

Richard Pyle (2008) provides the following example that I will reformulate slightly and summarize somewhat to fit my purposes.[50] Though on the long side, this example illustrates how complicated the correlation between species names and concepts can be as part of practice, and will help clarify some of the finer points made from this point forward. It will serve as an example that I can point to when describing particular issues related to name complexity.

Imagine two hypothetical species of fish are extracted from the a pool of water in the wild believed to be part of the same genus, named as:

Fish 1: *Holocanthus fisheri* (Barnthouse 1904) sec. Barnthouse 1904[51]
Fish 2: *Holocanthus acanthops* (Subramanian 1922) sec. Subramanian 1922

Then Myers comes along and decides that *Holocanthus fisheri* is actually part of another genus, *Xiphypops*, so he renames it with a new combination moving the genus:

Fish 1: *Xiphypops fisheri* (Barnthouse 1904) sec Myers 1933
= *Holocanthus fisheri* (Barnthouse 1904) sec. Barnthouse 1904

But notice that the concept hasn't changed; it has the exact same circumscription (description) as Barnthouse 1904.

Then imagine a third scientist described another fish from that same pool, and describes the following as part of a new genus:

Fish 3: *Centropyge flavicauda* (Van Winkle 1933) sec Van Winkle 1933

But Van Winkle also thinks that *all of the fish* from this pool are from this same *new* genus, so she decides to move *all* of the others into the same genus as well:

[50] The scientific species names and years used by Richard Pyle in his original PowerPoint presentation have been retained, however, the author designations have been changed.

[51] Where the italicized name is the valid scientific name (genus and species) that a scientist, Barnthouse, described and published in 1904. *Sensu* (often abbreviated *sec*) is a Latin term that means "in the sense of" and is often used at the end of names to indicate that the author used the species concept "in the sense of" whoever is cited. Thus, above, since Barnthouse described the species for the first time in 1904, it is *also* Barnthouse's concept.

Fish 1: *Centropyge fisheri* (Barnthouse 1904) sec Van Winkle 1933
= *Xiphypops fisheri* (Barnthouse 1904) sec Myers 1933
= *Holocanthus fisheri* (Barnthouse 1904) sec. Barnthouse 1904
Fish 2: *Centropyge acanthops* (Subramanian 1922) sec. Van Winkle 1933
= *Holocanthus acanthops* (Subramanian 1922) sec. Subramanian 1922

Pyle continues to describe how yet another individual comes along and decides that Fish 3 is actually a synonym of Fish 1, and proceeds to bring those two species groups under one genus:

Fish 1 and Fish 3: Centropyge fisheri (Barnthouse 1904) sec. Brown 2003
> *Fish 1: Holocanthus fisheri* (Barnthouse 1904) sec. Barnthouse 1904
> *Fish 1: Xiphypops fisheri* (Barnthouse 1904) sec Myers 1933
> *Fish 1: Centropyge fisheri* (Barnthouse 1904) sec Van Winkle 1933
> *Fish 3:Centropyge flavicauda* (Van Winkle 1933) sec Van Winkle 1933
= *Fish 1: Centropyge fisheri* (Barnthouse 1904) sec Van Winkle 1933
+Fish 3: *Centropyge flavicauda* (Van Winkle 1933) sec Van Winkle 1933[52]

And herein lies the problem, according to Pyle, "so now [*Centropyge fisheri* (Barnthouse 1904) sec. Brown 2003] applies to this whole circumscription of organisms [the species groups that originally represented the populations of Fish 1 and Fish 3]. So it's a different concept using the same name…that's the problem we have … names don't match perfectly to concepts. Sometimes the same concept goes by different *legitimate* names and sometimes the same name can refer to different *legitimate* concepts" (my emphasis) (2008). Taxonomist Nico Franz, et. al. highlight this disjoint between "names and taxonomy" and the problems it causes when a "name and its meaning evolve independently" (N. Franz, Peet, & Weakley, 2008, p. 64). As is seen above, the token for any given species is entirely dependent on the metadata attached to the name itself—information that is, at best, occasionally included as part of a name as it circulates within technical systems. But in order for a system to work, congruence between names and concepts must be solidified in some mechanized way. Expanding on this issue in particular for a moment, please refer to Figure 11 below, which is from David Remsen's article, "The use and limits of scientific names in biological informatics" (2016). He notes how there is no *direct* relationship

[52] Where ">" means "the synonym of," and "+" indicates that the two species concepts melded coming together to form one larger concept (in this case, that of *Centropyge fisheri* (Barnthouse 1904) sec. Brown 2003).

between the symbol (the name) and the referent (the real world object), which is indicated by a dotted line.

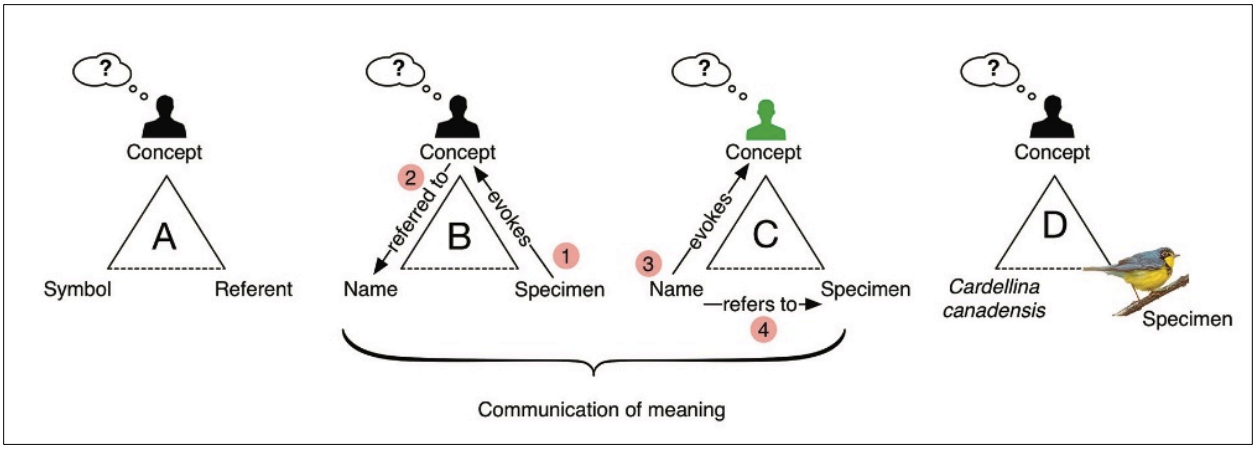


Figure 11. The Semiotic Triangle. "The semiotic triangle describes how names communicate meaning" (D. Remsen, 2016, p. 211).

In this article Remsen draws our attention to two "sub-domain of semiotics, *semantics* ("the relationship between signs and things…their meaning") and *syntactics* ("relation among signs or symbols within formal structures") (2016, p. 210). Remsen makes the argument that taxonomy has two essential concepts that mirror this distinction: that of taxonomy (semantics) and nomenclature (syntactic) (pp. 210–211). Names work in formal systems (in this case, technical systems), while taxonomy is concerned with the relationships between the described *concept* of the species and its nomenclatural representation. The former will be discussed in more detail below, but here I'll focus on the semantic relationship between word and concept. In Figure 11, Section A, Remsen illustrates the basic Piercian "triadic relation" (Short, 2007, p. 30): the sign, the object, and the interpretant. The goal of the taxonomist is to interrelate these concepts as tightly as possible to facilitate ready communication of concepts. Nomenclature (the sign or sign-bearing) must, through the careful and diligent description (the referential extensions" (N. M. Franz et al., 2016, p. 646)) of a type specimen (the object), "translate" (Atkin,

2013), as accurately as possible, the "mental equivalent" of the concept as it was originally conceived to another scientist (the interpretant) (see Section B, Figure 11). The interpretant must then be able to work in the opposite direction: looking at the name, as well as its associated circumscription, the interpretant must be able to conceive of the original species concept as close as possible to as it was imagined by the original describing taxonomist (see Section C, Figure 11).

This "communication is facilitated," however, "only when the concepts (not the objects) are approximately congruent" (D. Remsen, 2016, p. 211). This is no easy task, especially when the "cardinality between syntax and semantics … is one-to-many" (2016, p. 211), meaning that any one name can reference an innumerable set of valid (and invalid) species concepts. Thinking about Pyle's fish example above, the *articulation* of a species concept is far more than *one* document and one type. In the simplest case, it may be that one type specimen and one publication are all that is necessary to reconstitute a concept. Things can get far more complicated, however, when multiple type specimens are taken into account in the production of a species circumscription (a holotype, along with a collection of paratypes, for example).[53] Which says nothing of the emergent complexity that is produced when names and concepts develop over time; in the case of the *Centropyge fisheri* (Barnthouse 1904) sec. Brown 2003, the newly created genus and species taxon concepts constitute *at least* five publishing and

[53] A holotype is the "single specimen used by an author, either the only specimen he found or one of several found, but the only one designated as a type" (Winston, 1999, p. 104). A paratype, on the other hand, "are specimens that the person making the original material examined while carrying out the work" (1999, p. 104)—so when a group of organisms are examined to establish a species concepts, the paratypes are those types that are *not* the holotype.

nomenclatural acts and potentially dozens of types specimens that were used in all of their original genus and species descriptions.[54]

Infrastructures such as the Catalogue of Life are designed to facilitate the conveyance of concepts by controlling nomenclature. As Paul Kirk, Senior Biosystematist and Mycologist at Royal Botanic Gardens, Kew, conveyed, "Stability [is] the single most important thing about nomenclature: 'this name calls out the provision of a stable method of naming taxa avoiding ambiguity and confusion,'" but stability is contingent on a number of factors (2016). "Customers don't like name changes ... medical mycologists don't like name changes...governments, [IUCN] Red List people ... every customer of the products of taxonomy with the nomenclature behind it don't like name changes. Unfortunately, taxonomists like changing names…." (2016). The Catalogue, and nomenclaturalists in general, go through a great deal of steps and processes to document all of these name changes over time in order to provide ready and (as) accurate (as possible) access to names data. The next section will document the process of *controlling* names within technical structures and professional standards.

Part II: Nomenclature: Toward the Appraisal of Knowledge

From bio-documentary description to exploitation.

Managing names requires a system of control. One of the fundamental contributions of Patrick Wilson's *Two Kinds of Power* (Wilson, 1968) is his articulation of exploitative power in addition to the already-familiar notion of descriptive power as it applies to bibliographic practice. As I will illustrate below, the production and management of names within the Catalogue of Life can be seen as exemplifying a spectrum between these two theoretically distinct phenomena: as

[54] See also Nico Franz's article, "On the Use of Taxonomic Concepts in Support of Biodiversity Research and Taxonomy," for an in depth discussion on the evolution of taxonomic perspectives (2008, pp. 63–65).

names move from pools of undifferentiated lists and nomenclatures, toward more structured and taxonomic formats, they become linked in knowledge discourses that are created to allow for more effective retrieval and user-oriented mediating interfaces. Further, the imposition of taxonomic frameworks in biodiversity studies exemplifies how descriptive and exploitative frameworks cannot, in practice, be seen as distinctly separate intellectual activities. I will first briefly situate the concepts of descriptive and exploitative power as articulated by Wilson then move toward a concrete example of how such concepts are at play in the Catalogue.

Descriptive power, as described by Wilson, is "not a very adequate term for an ability to line up a population of writings in an arbitrary order, to make the population march to one's command. The wielder of perfect bibliographic descriptive control can have summoned up every writing that fits his arbitrary description, so long as the applicability of that description to particular writings can be discovered without any consideration of virtues or vices" (1968, p. 25). What Wilson calls an "evaluative neutral" approach is what others have identified as the "traditional" descriptive practice of cataloguing rooted within the field of librarianship (Coyle, 2016, p. 39; Morris & Van der Veer Martens, 2009, p. 224). In particular, this approach takes textual entities (documents broadly construed), and as indicated by Svenonius, "requires that bibliographic descriptions be constructed to reflect the way bibliographic entities represent themselves" (2009, p. 71). The attributes, terms, subjects, etc., applied to any given object can then be used to collocate disparate resources; Wilson uses the examples of wanting to bring "all of those writing by Hobbes, all those discussing the doctrine of eternal recurrence, all those containing the word 'fatuity' (1968, p. 22). This concept has had considerable impact within classification and knowledge organization research, especially since its primary aim is "to provide a complete listing of all members of a class" (Morris & Van der Veer Martens, 2009, p.

224). But the act of descriptive power can only take a system so far. Optimally, the task of information control within documentary spaces is *also* to help individuals retrieve those documents that are most suited to their individual needs. Wilson articulated exploitative control to satisfy this expressed need as a function of bibliographic systems.

Exploitative control is "the ability to make the best use of a body of writings … the wielder of perfect exploitative control has merely to say what he wants writings *for* and is then provided with what will suit the purpose best, whatever it is" (Wilson, 1968, p. 25). If descriptive power was the intellectual act of creating bibliographical representational structures, exploitative control allows that power to be put into practice within *systems* in ways that meet individual needs. Scholars have also made the reasonable leap to say that exploitative control is a way in which we can think about information retrieval within the area of information science (Coyle, 2016, p. 40; Smiraglia, 2007). That *one* can "have the power to procure the best textual means to ones end" (1968, p. 22, my emphasis) is a vital contribution by Wilson to Information Studies, not the least of which because it brings a focus onto the user and has helped refine the ways in which we conceptualize and sharpen our information systems for specific uses. That is to say: information and knowledge organization systems are made for communities to *use* information that best suit their information needs. Finding sets of documents is one thing, but it is quite another to say that said information is *useful* as a function of its contextualized use. Karen Coyle articulates the importance of this concept within the discipline of bibliography:

> [Wilson] begins by stating something that seems obvious but is also generally missing from cataloging theory, which is that people read for a purpose, and that they come to the library looking for the best text (Wilson limits his argument to texts) for their purpose. This user need was not included in Cutter's description of the catalog as an "efficient instrument." By Wilson's definition, Cutter (and the international principles that followed) dealt only with one catalog function: "bibliographic control" the second is the appraisal of texts, which facilitates the exploitation of the texts by the reader. This has traditionally been limited to the realm of scholarly bibliography or of "recommender" services (2016, p. 39).

Richard Smiraglia also notes that the concept of exploitative power "gave succeeding generations of researchers a means of measuring efficacy of systems for knowledge organization. Whatever enabled exploitative power was efficacious; whatever obfuscated exploitative power, and this was most of the bibliographical apparatus, was not efficacious" (2007, p. 1). Retrieval is the "operationalization" (Tennis, 2006) of purpose predicated on the extensive and effective establishment and implementation of descriptive structures. And indeed, as Wilson makes quite clear, while the "two sorts of power have been contrasted as sharply as possible, as is desirable in an exercise in analysis," "no doubt the limited power actually possessed by people are complex mixtures of the two" (Wilson, 1968, p. 29). The process of nomenclatural control is not different in kind from that of documentary retrieval. The ultimate goal is to call up a name, and all of its associated conceptual documentation, and to *use* that information for scientific and taxonomic work.

Systematizing nomenclature.

So then, how can we frame the concepts of description and exploitation within the Catalogue of Life? By examining the process of nomenclatural management within this system, I can illustrate how these two principles play-out as part of the practice of building authoritative databases. In November of 2016, the Catalogue of Life invited me to a closed workshop titled, "NAMES in November" (Global Biodiversity and Information Facility, 2016b; Global Names Architecture, 2016b; Pape [@fleshflies], 2016) hosted by both the Catalogue of Life and the Global Biodiversity Information Facility (GBIF) in Leiden, Netherlands. The function of this gathering was to "to discuss classification, taxonomy, nomenclature, and vernaculars in context of the practical aspects to develop a single fully open-access and 'all encompassing' taxonomic backbone maintained and owned by the taxonomic community that will serve the needs of the

broader user community" (Schalk, 2016a). Even while the Catalogue, GBIF, and other participants of the iLife consortium have certainly made incredible strides in coordinating nomenclatural and taxonomic work, a concerted and clear effort needed to be made to refine service interoperability based upon this vision of a single, sustainable taxonomic backbone, and to be able to maintain this structure into the future. As previously discussed, names are a foundational part of not only taxonomic work, but biodiversity work in general. Without an adequate pool of *authoritative* and *controlled* set names, accurate scientific work cannot proceed adequately, and the proliferation of data sources compounds the generalized problem (Boyle et al., 2013). As Brian Heidorn has indicated, huge volumes of data, the "primary outputs of the scientific enterprise," are still unavailable to scientists, meaning they are "dark data" and thus unavailable, or invisible, to the scientific community for the production of new knowledge (2008, p. 280). The first order of business for taxonomic databases is to get *all possible names* aggregated so that taxonomic authorities have the basic building blocks available to them. In order for biodiversity *information* to become semantically meaningful and truthful (Floridi, 2010, p. 20) all the generalized data needs to be consulted at the outset.

Given these non-trivial issues, partitioning and coordinating the management of names is a top priority for these biodiversity scientists and informaticians. One fact was clear about the NAMES meeting: the problems posed, and the solutions articulated, were to remain focused on the intellectual, theoretical, and policy needs to satisfactorily coordinate knowledge, and *not*—as usually seems to be the case—to concern itself with "technical solutions" or computational capacities. The technical issues were to be handled at a completely different meeting currently unscheduled within the 2017 year.[55] What follows are the salient issues that were raised by this

[55] The outcome of this NAMES meeting will be a position paper that delineates the general impediments to nomenclatural and taxonomic coordination.

group of professionals, as well as some preliminary schematics for how the nomenclatural landscape can be adequately managed given the various temporal and professional complexities related to species concepts that were described above.

In order to gain a sense of the scope of the meeting, here are some select questions the meeting was intended to ponder:

- What does the world need as a nomenclatural resource?
- How can we (sustainably) get the content we need?
- How can we build on existing databases and expertise?
- What models may work for curation of these data?
- Should crowdsourcing be used some form?
- How should we handle newly-published names? A central repository?
- What about vernacular names, manuscript or informal names ("Carabus sp. Leiden"), Barcode Identification Numbers, etc.?
- How do we deal with conflicts?
- Who should be allowed to contribute, and how should these rights be partitioned across datasets?
- How do we credit contributors?
- How to deal with (living) data quality indicators?
- What need is there for different national, regional or thematic views?

As can be seen, many of the questions relate to our previous discussion of contingent documentation. Two primary documentary issues at hand are: (1) how this ecology is going to delineate clear boundaries between one entity and another given similar and overlapping nomenclatural needs within each organization, and (2) how will the Catalogue of Life, if indeed it were to continue as the central nomenclator and resource for validated taxa, manage its porous boundaries given the necessity of ongoing contributions, necessary re-combinatory views (national, regional, thematic), and (though the subject of a later chapter) the increasing production of barcodes that are not "meaningful" within the system (in the sense in which Floridi discusses above with regard to information).

In Figure 12, below, you will see the general parameters (and their overlap) between all the potential names (the entire namespace) and the Catalogue of Life (with structured names into

taxonomic knowledge) as articulated by Thierry Bourgoin.[56] This schematic illustrates a birds-eye view of the stages and steps of nomenclatural (and taxonomic) management for scientific purposes, as well as the intellectual and editorial spectrum by which all undifferentiated name text strings become organized into valid taxa and become associated with valid concepts. Dr. Bourgoin outlines five general actions within the biodiversity knowledge continuum that merit addressing: Stage 1: Names as Strings (all names); Stage 2: Scientific names; Stage 3: Potential taxa/Chresonym (Names+Reference+Usage); Stage 4: Valid Taxa; and Stage 5: Alternative Classifications and Phylogenies. Note the orange circle labeled "Management classif." toward the right of the illustration—this is where the Catalogue of Life management classification fits within this schematic. For the purposes of this chapter we will only concern ourselves with the procedural space to the *left* of the production of valid taxa and the management classification, or Stage 1 through Stage 3.[57]

[56] It should be noted that the terminology used in this figure conforms to the *International Code of Nomenclature for algae, fungi, and plants*, which is clear by Bourgoin's usage of basionym, which is defined in Article 6.10 of the ICNAFP (International Association for Plant Taxnomy, 2011). The equivalent in the ICZN is "original combination."

[57] Taxonomic production and the Catalogue's management classification and its multifold issues will be the subject of our next chapter after the nomenclature landscape has been adequately described.

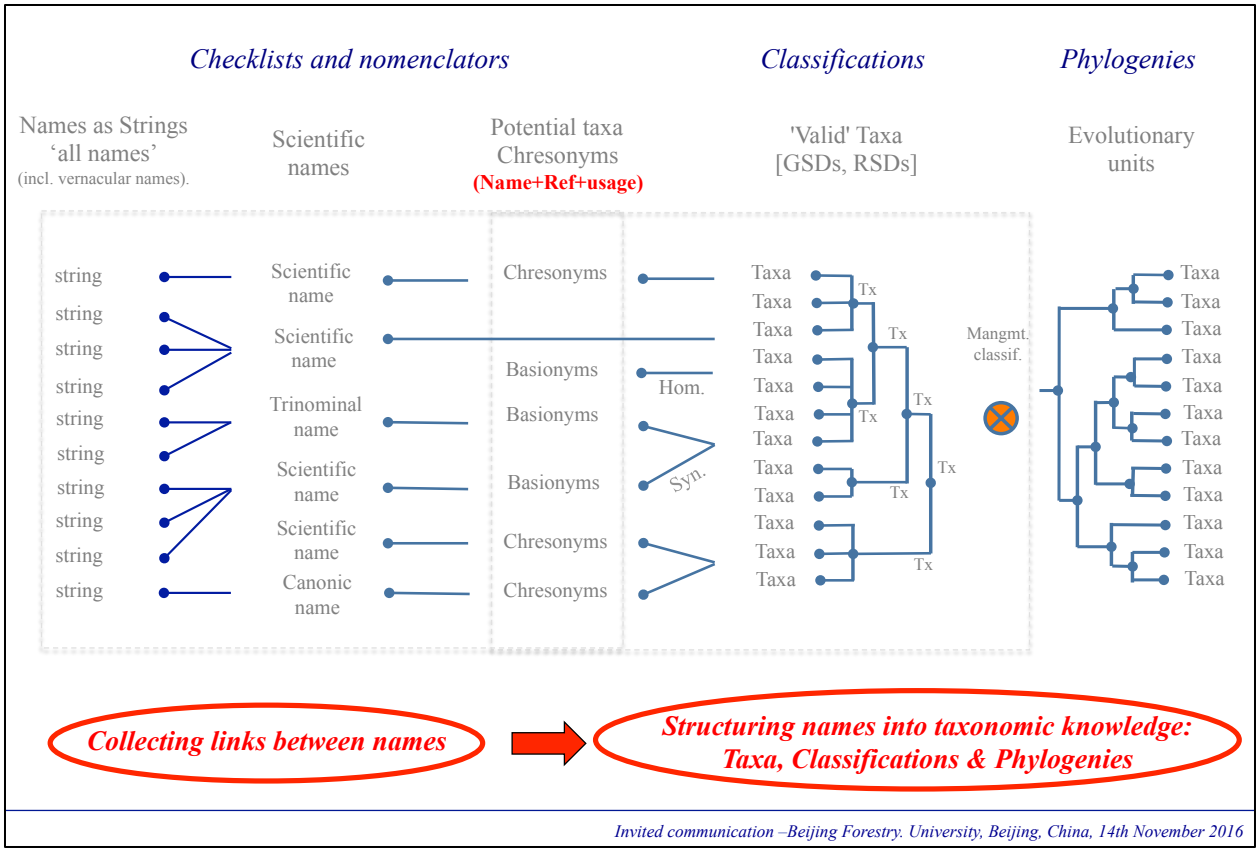


Figure 12. Global Names-Catalogue of Life Parameters. Source: Thierry Bourgoin (2016).

Toward a universe of all possible tokens: Global Names Index to Nomenclators

As we have seen, names represent complex concepts, and therefore they represent an amalgam of material documentation and intellectual processes that are necessary in order for them to function as effective representational instruments within the biodiversity arena. Maximum ability to exploit the biodiversity documentary universe is contingent upon on the ability for scientists to have ready access to the information, data, and knowledge produced as part of their work, hinged as it is upon the *names* that organize this disparate content (in whatever form that information, data, and knowledge might take). As important as it was for Patrick Wilson in *Two Kinds of Power* to describe, and subsequently exploit, texts, one of the necessary contingencies to maximize both kinds of power was that all of the available texts within the bibliographic universe needed to be *accessible*. The case is no different in the

biodiversity world than in the bibliographical, though in the former's case, names become the extensive agent and portal by and through which database environments can point to and access any particular document within the biodiversity domain. Fundamentally, the *nomenclatural* act is both a representational object as well as a descriptive one. In the bibliographic world, subject terms, facets, controlled vocabularies, etc., all serve a collocative function, but in the biodiversity world, this function rests almost completely on the name itself.

"Names as String"/'all names,' the first stage in Bourgoin's schematic, is the entire hypothetical pool of undifferentiated names produced around the world—those in scientific form, vernacular/common form, etc. Due to their un-validated status, names in this space are text strings of *potential* significance. I mean significance here in two sense: (1) in terms of *well-formed* units as that will function with a computational ecology of specific name formulations, and (2) in terms of whether or not it is the *valid* use of the name. The Global Names Architecture (GNA) (Global Names Architecture, 2017c) has arisen to serve a vital name-differentiating function in the biodiversity world. The Global Names Architecture has been implemented as "a system of web-services which helps people … register, find, index, check and organize biological scientific names and interconnect on-line information about species" (2017b). It should be noted that the GNA's strength lies in names that derive from the zoological or botanical codes of nomenclature, so as has been the case, my examples and general statements refer to the rules that manage the production of names within these two broad domains. The name-space within the GNA that represents the entirety *of all possible* name instances is called the Global Names Index (GNI). The GNI is "an index of name strings in the broadest sense (including code-compliant scientific names, vernacular names, surrogates, identifiers and erroneous versions of names), with links to sources that have the names and to the associated

data or metadata associated with names. GNI is a core element of the GNA. This is referred to as a 'dirty bucket' because it is a raw list; but every item in it has a scientific context or indexes scientific content" (2017d). Content for the GNI is contributed by many organizations and collected from online sources (Global Names Architecture, 2017a). As was previously indicated, the Biodiversity Heritage Library (BHL) is an enterprise that has cropped up to manage the digitization of historical documents in order to contribute names to the global names-collecting enterprise, and is currently attempting to coordinate the contribution of their contextualized name data to the GNI space. Libraries including the Natural History Museum, London, and the Smithsonian National Museum of Natural History in Washington, D.C., contribute digitized literature that are then made available to the global biodiversity community as a central source for historical taxonomic literature. As names are collected with the assistance of optical character reader (OCR) software, a good deal of errors are introduced to the text, which is another task facing those differentiating this namespace.

As of August 2015, the GNI data bank contained upwards of twenty-three million names strings which contains a mixture of both "good" and "bad" names (Global Names Architecture, 2016a)—a combination of current, outdated, valid, and invalid scientific names. The process of assessing this namespace, orthographic format, adherence to nomenclatural codes, and, ultimately, getting "[endorsement] by a taxonomist" brings us to Stage 2 of the GN-CoL Parameters scheme produced by Bourgoin: that of assessing which of these name strings qualify as *well-formed* scientific names.[58] Given the Catalogue's functional estimate of 2 million total global species, actions need to be taken to differentiate what should be allowed to proceed in

[58] Dr. Bourgoin, aside from scientific name, notes that Stage 2 also includes the aggregation of canonic and trinomial names. Canonic names are names stripped of any other information aside from their Latinate components, while trinomials are formed when "when a subgenus or a subspecies is added to the species name" (Global Names Architecture, 2017d).

their checklist system and what is deemed "bad" (invalid, not agreed-upon) name data (See also Pyle, 2016, pp. 270–271).[59] Thus, after all name strings have been aggregated, the next step is to decide which are correctly formed scientific names in this sense.[60] The GNA provides an overview of what qualifies a term as being "scientific":

> Scientific names are latinized [*sic*]. Alternative names are common names (also referred to as vernacular or colloquial or familiar or informal - such as cat, dog, crow, maple); there are surrogates for names (being name-strings that may refer to a culture, or some term by which an organism is widely known in, for example, research settings) … A species name is distinctive. It is made of two parts, a genus and a species part (Homo sapiens, Drosophila melanogaster). These are binomial names. Not all latinized [*sic*] binomial names that have a capital letter at the front are organisms: 'Anorexia nervosa is an eating disorder, and Habeas corpus is a legal term') (2016a).

Lists endeavoring to collect all names, across all kingdoms of the biodiversity world, must contend with hundreds, perhaps thousands of database index lists, many of which are far less curated than the examples we'll cite from Kew below. On top of these many data sources, each list contains potentially thousands or hundreds of thousands of names, which quickly makes for a rather onerous and messy process of disambiguation and error reconciliation for the governing body/bodies that attempt to take on this management role. Both the distributed nature of these various database sources, and the multiple formats in which these lists exist, present multifold issues for nomenclatural organization. Richard Pyle articulates the main issues associated with names within the computational arena, especially with regard to digitizing analog material:

---

[59] Even if we take a more generous approach to the global species count, estimating upwards of 8.7 million species (Mora, Tittensor, Adl, Simpson, & Worm, 2011), the GNI name pool is still comparatively large.

[60] Bourgoin's model presupposes a nomenclature pool that contains *only* scientific names. Discussions were had during the NAMES conference concerning whether or not the Catalogue should also include common names and various other non-formal iterations of the valid scientific version. While the CoL contains common names as part of its data set, the question becomes whether or not *comprehensive* coverage of common names should be a general aim of their efforts. NAMES participants voiced support for both opinions: those that articulated a need for comprehensive common names pointed to their importance as mechanisms for retrieval; those that preferred common names as optimal, but not priority, pointed to the pure *volume* of possible common name instances. One issue here is that common names—produced as they are in local contexts—can be expressed in a number of languages, so any given scientific name produces any number of common forms. Scientific names are more manageable in this context because language is not an issue since they need to be expressed in Latinate formulations regardless of the country in which they originated.

> The rapid evolution in recent decades of computer database management software, and of information dissemination via the Internet, have both dramatically improved the potential for streamlining the entire taxonomic process. Unfortunately, the potential still largely exceeds the reality. The vast majority of taxonomic information is either not yet digitized, or digitized in a form that does not allow direct and easy access. Moreover, the information that is easily accessed in digital form is not yet seamlessly interconnected in an effort to bring reality closer to potential, a loose affiliation of major taxonomic resources, including GBIF, the Encyclopedia of Life, NBII, Catalog of Life (*sic*), ITIS, IPNI, ICZN, Index Fungorum, and many others have been crafting a "Global Names Architecture" (GNA) (2016, p. 261).

Bourgoin's first stage is, as mentioned previously, primarily orthographical, ensuring that the GNI pool is pared-down to terminology that generally match the spelling, language, and formal conventions of the scientific nomenclature system. To perform these kinds of text differentiating activities, however—especially when dealing with data sets with over twenty million text strings—is a nearly impossible task for any group of individuals to achieve, particularly since new names are being constantly added into the GNI. Name matching services, based on various lexical algorithmic software, are available to help facilitate this process (Pilsk, Kalfatovic, & Richard, 2016; Vanden Berghe et al., 2015). GBIF has its own open source name parsing tool that "can be used to automate some processes while digitizing or curating lists of scientific names" (GBIF, 2011, 2017). GNA also has its own Global Names Resolver (2017b) that examines text strings to assess whether a name is a scientific name, correctly spelled, currently in use, and a host of other variables (2017a). But algorithms and name resolvers have their limitations, especially given the name-to-concept mapping issues that were described above. Minute differences between nomenclatural codes can also present some issues for automated resolvers, since orthographic conventions between the botanical and zoological codes sometimes call for different (and conflicting) nomenclatural conditions that must be assessed individually to come to a decision. One example, noted by Richard Pyle (R. L. Pyle, 2008, n. 11:53) is the "ex" designator in a name such as *Anthias ventralis* Randall 1979 ex Thompson, which in the zoological code indicates that 'Randall 1979 published a new name *Anthias ventralis*, "but based it on the intellectual work of Thompson"' (2008). In the botanical code, however, Pyle notes the

same concept is conveyed in the exact opposite manner and order as dictated in the zoological code: '*Anthias ventralis* Thomp. Ex. Rand.' Such inconsistencies plague the nomenclatural landscape and illustrate the extent to which names-as-tokens are less than optimal and consistent.

The next step in the process, as Patterson, et. al, have articulated, is that scientific names must also be "compliant with the relevant code of nomenclature, or, if the codes do not apply to them (for example, because they are names of high ranking taxa), they are written in a comparable form consistent with the expectations of biologists" (2016, p. 3).This is the task of nomenclators. Nomenclators provide a listing of "code governed facts" (T. Orrell, private communication, February 6, 2017)—the sandbox space where well-formed and scientifically-valid names can be aggregated for subsequent taxonomic use. As conveyed by David Patterson, et al., nomenclators must first "indicate the correct orthography of each scientific name, [then] accompany it with the name of its author, the date when the name was introduced, and a citation pointing to where it was first used. This may be in the form of a condensed micro-citation…Nomenclators develop lists of scientific names of taxa, but are not lists of taxa because a nomenclator makes no evaluation as to the taxonomic status of a name" (2014, p. 2). Nomenclators and listings such as International Plant Names Index (IPNI), located at Royal Botanic Gardens, Kew, live in this name and text space. Lists serve a variety of "requirements" (Croft et al., 1999, p. 320) within the biodiversity world, and ultimately must meet the following needs:

- they contain names at each rank with their places and dates of publications,
- they employ consistently community standards for data and abbreviations, etc.,
- they represent authoritative expert knowledge,
- they are exhaustive,
- they are kept current,
- they are freely accessible,
- they can be queried in a variety of ways, with any needed information being downloadable,
- and they entail minimum effort and cost to the systematic community for their upkeep (1999, p. 320).

Many of these articulated requirements, when inserted in distributed infrastructures, become problematic to control for any number of reasons, including apportioning responsibility for these activities within the iLife platforms; being able to gather *all* names in some unified fashion' managing communication of feedback between platforms to maintain database currency; establishing long-term cooperative funding, etc. Some of these issues formed the core questions and concerns posed to the NAMES in November group in Leiden.

Nomenclatural listings are generally maintained as *databases* in contemporary practice, especially if one wants to remain relevant and share information beyond local labs or sites. Lists dating back to the 18th century, however, are equally valid for today's research given how important historical knowledge is to the contemporary taxonomic community. Into the early 1980's analog lists were the most prevalent and only form of data used in taxonomic practice (recall here the brief discussion of data types and data siloes in the opening chapter of this manuscript). Once nomenclator listings started making their way into the digital arena, all of these legacy documents needed to be converted into digital, tabular form. Index Fungorum (Royal Botanic Gardens, Kew, 2017c), an initiative that initially began at the Centre for Agriculture and Biosciences International (CABI), is an example of such retroactive conversion. While interviewing Paul Kirk, current editor of Index Fungorum, at his Kew office in London, he began our meeting by pointing to a row of books sitting on the shelf above his desk that collectively formed the core text for the now-digital Index Fungorum. When the Mycology Institute integrated information technologies into their work processes, Paul Kirk and colleagues meticulously transferred the *Index of Fungi* (Petrak, 1969), a 625-page dictionary of fungal terms, as well as numerous index cards inherited from subscription-based index listing services prior to the 1980's, into digital form in the late 1980's and 1990's (Kirk, 2016). These now

aggregated print sources form the historical core of Index Fungorum, which as of April 2017, holds 532,288 online records contributed by over 1,000 individual authors (2017b). As is evidenced by Index Fungorum, lists are usually created for specific domains (for plants, fungi, lichen, beetles, and other species a biodiversity specialist may emphasize and records as an organic result of working practices) for use at some local and institutional level. The Royal Botanic Gardens, Kew's International Plants Name Index (IPNI) (Royal Botanic Gardens, Kew, 2017e) is another prevalent example, the composition of which is bounded, generally, by those organisms governed by the International Code of Nomenclature for algae, fungi, and plants.

Building networks: Linking tokens and documents.

I would now like to more closely examine the linking activity that is undertaken by nomenclators, essential as it is in connected various forms of biodiversity evidence. In reality, the orthographic cleanup described above happens in tandem with this step, but for illustrative and analytic purposes it makes sense to discuss them as distinct activities. The GNA provides a mechanism for this kind of linking activity, a service called the Global Name Usage Bank (GNUB). The purpose of GNUB "is to index and assign persistent globally unique identifiers (GUIDs) to Agents, References, and Taxon Name Usage (TNU) instances (among other relevant data objects). Agents are people and organizations, and in the context of GNUB mostly represent Authors of References. References include all published literature, as well as many forms of unpublished documentation (e.g., unpublished reports and manuscripts, specimen labels, herbarium sheets, field notes, etc.)" (R. Pyle, 2016, pp. 270–271). In other words, the GNUB documents all *usages* of names so that they become part of an interconnected network of collective biodiversity meaning. Further, the TNU might also be associated with any number of metadata elements relevant to the nomenclature process (or part of the species concept and

taxonomic articulation process), such as type designation (which may include a set of specimens—holotypes, paratypes, syntypes, etc.—from which the circumscription was articulated), geo-coordinate estimations, molecular data, figures, etc.[61]

Recall that *any* name must be validly published in order to be deemed in circulation as *known* scientific information, and the specific date in which a name is coined is used to assess the priority of one naming act over another. It is one thing to list well-formed names, but it is quite another to be able to give those names context and conceptual heft by indicating how that name has been invoked as part of the discourse of biodiversity knowledge. The validation and hierarchy of term usages cannot be completed without understanding the "infinite network" (Deleuze, 1987, p. 587) of signs and how these signs refer to the objects that give them credibility as part of the rules that govern nomenclature. Returning once again to Figure 12, there are two faint boxes surrounding Stages 1-3 of the diagram, as well as Stages 3-4—the box to the left is the domain of entities like the GNI, GNA, and nomenclators (entities that govern the more objective assessment of valid "factual" forms of names), while the box on the right is the domain of the global species databases, regional species databases, and the Catalogue of Life (where more subjective assessments about taxa are made). The center area, Stage 3, where linkages between "Names+References+usages" occurs, is the start of where "knowledge" begins to play a role in nomenclatural spaces. Note also the red arrow and red circles running at the bottom of Bourgoin's figure, and how we are slowly moving away from organizing entities according to prescribes rules to more subjective modes of knowledge-making. As the GNA articulates, "All such usages [of the name statement] of all species in all documents make up humanity's knowledge of the biosphere" (2015). To be part of collective knowledge in this sense, terms

[61] Syntypes are "two or more specimens selected from the available material to serve as types" (Winston, 1999, p. 104) when no holotype is designated.

themselves must be semantically meaningful, according to Floridi, in that names are given meanings that allow them to circulate as defined components within the nomenclatural system. In order for any taxonomic assessment to take place during the latter part of this process, all of the basic tokens must be properly linked to the descriptions (circumscriptions) of the concepts they represent, so they can then be recombined and altered relatively easily moving forward.

Calling out two specific issues in mapping historical name usages in Stage 3, Bourgoin identifies the assessment of basionyms and chresonyms as particular kinds of nomenclatural complications that occur in this indexing space. Disambiguating basionym and chresonym use requires more expert taxonomic knowledge to categorize within the name space (certainly other issues present themselves as part of this process, though given space constraints, I will only be focusing on those mentioned by Bourgoin). In the first instance, chresonyms are "references to the use of a name. They can be presented in many formats … Problems arise when the format is simply "*Name* + user" (such as *Homo sapiens* Smith, 2005). This is intended to indicate Smith's use [not designation of the original concept] of *Homo sapiens* in an item published in 2005. This form is not distinguishable in form from code-compliant names" (David J. Patterson et al., 2006, p. 371) (which follows the form, *Name* + species concept designator). The problem in this case is that, without careful attention, these names can often be mistaken as a synonym for a validly published name, but this is not the case, as a chresonym is not code-compliant (Global Names Architecture, 2017d). Such instances require "the need for taxonomic intelligence to disambiguate them from homonyms" (David J. Patterson et al., 2006, p. 371) (an instance of two names identically spelled representing two different taxon).

Given the continual updating of the relationships between species concepts and the names that represent them, close attention must be paid to how these two evolve in tandem over time.

Leading to the second complication, that of the basionym,[62] which is a linkage between name forms that indicate "the relationship between a new combination and its original combination …really just a pointer back to the original combination of a name" (R. L. Pyle, 2008). The basionym often forms part of the new name combination in some way. For example, *Micromussa amakusensis* Veron, 1990 (World Register of Marine Mammals (WoRMS), 2015) is the current scientific name for a species of marine coral most frequently found around Japan, but the previous name, *Acanthastrea amakusensis* Veron, 1990, is the *previous combination* of species when it was re-classified in the *Acanthastrea* genus and removed from the *Micromussa* genus. Thus, mapping out basionym designations within a nomenclator refines the networks of semantic units. Closely mapping out transformation of name forms over time, spurred on by changing interpretation of species circumscriptions, makes for more accurate communication of scientific knowledge. The NAMES meeting made it very clear that combinations and names usages were of high priority in an idealized nomenclator system.

Finally, the homotypic synonym (for the botanical code)—or objective synonym (for the zoological code)—refers to two names for the same taxon that share a type specimen(s) as well as a basionym (Global Names Architecture, 2017d). As Patterson et al. have indicated, "The most significant known challenge with the use of names as metadata is the 'many names for one taxon' problem" (2016, p. 3). In all of these cases, the purpose is not merely to connect the *names* within these index spaces, but also to connect the various circumscriptions, references, and historical variations to facilitate the accurate and efficient application of taxonomic structures using the metadata surrounding these name tokens.

[62] Or original combination as it is expressed in the zoological code (Rapini, 2014).

On another level, synonyms are essential tools in making an information system *exploitable* in the sense that Patrick Wilson uses the term. Thomas Orrell of the Smithsonian National Museum of Natural History made the point, on numerous occasions, how important "synonymic amplification" was for the discovery of information (T. Orrell, personal communication, June 15, 2016). The more synonyms that are mapped and documented within the system, the more likely an individual will find the information needed. "Relationships are at the very heart of knowledge organization," as Rebecca Green has astutely noted (2008, p. 151), and in specific, interconcept relationships are of particular interest here. "With equivalence relationships, synonymy (e.g., dog, canine), quasi-synonymy (e.g., lexical relationships, paradigmatic relationships), and occasionally antonymy (e.g., good, evil) are expressed by choosing one of the set of terms as an authorized descriptor and using it in lieu of all others" (2008, p. 156). This description sounds quite familiar to the ways in which we can think about the relationships between a valid scientific name, its basionym, chresonym, and so forth. Gerald Guala (2016), also of the Smithsonian NMNH, examined the role of synonyms for searches in PLoS, PMC, PubMed or Scopus. Using scientific names from ITIS that were mapped to synonyms, Guala found significant increases in citation recall and precision rates when synonyms were used. Synonymic amplification, then, serves both a descriptive purpose within a nomenclator system (in that it allows scientists the ability to line up relevant documents), but it also provides an exploitative purpose (in that it also helps facilitate the best *use* of said information).

Before proceeding to Bourgoin's next stage of biodiversity taxonomic knowledge creation in the next chapters—that of articulating valid taxa, taxonomies, and, ultimately, phylogenies—it makes sense to stop and think about how all of the aforementioned name-token

control might be dealt with within the Catalogue and iLife's ecology. Granted, the NAMES meeting in Leiden was meant to focus on the theoretical aspects of nomenclature, but nonetheless, there was some discussion about how nomenclature might be handled procedurally and institutionally given its distributed nature and current challenges. In the end, this brief foray into technicalities allowed the participants to draw boundaries around what nomenclature issues they *should* and *should not* attempt to manage within each of their organizational boundaries.

Controlling complexity: The Catalogue of Life Plus.

One of the clear outcomes of the NAMES meeting was that there needed to be a focus on utilizing present infrastructures and existing informatics tools to streamline the process of nomenclatural control. The general consensus of participants was that the Catalogue of Life could manage this complex nomenclature activity given its already-deep investment in an edited, high-standard global checklist, its long-standing partnerships with global service databases, and its current initiative in progress known as Catalogue of Life Plus. Peter Schalk, Executive Secretary for the Catalogue of Life and Governing Board Chair for GBIF, explains the Catalogue's extension as follows:

> The Catalogue of Life has ten or eleven fields that we ask for from the [contributing] databases, which is fine if you, say, want to find out about the name, want to know the major synonyms (we don't have all of them), or want to know the hierarchy... then you can find what information you need. But there are people that want to have other, more-detailed information in those databases. Actually, the idea behind the Catalogue of Life is that you go to find a particular species and then you have the ability to drill down into the contributed databases where all the information is living. There is actually a demand for a Catalogue of Life that has a lot of information living at the bottom that is already indexed ... This kind of service would encompass a broader definition of the Catalogue of Life with more fields. Another thing is they would basically like to find *all* the names. Because you can't type in a name that we don't have because it hasn't yet been processed. But [taxonomists] do generally have an idea where it belongs—that's what GNA [has] done. They harvested all the names, threw [them] into a system, and made relations between the names so you know more or less where it belongs. But it hasn't yet been vetted by experts, and edited. That is where the Catalogue of Life Plus comes in. The idea is that you take the Catalogue of Life, and you create around it a cloud of all of the names that haven't been processed by the GSDs, but that can be linked to certain sectors. So when you do a search you can either say, hey, my name has been found, this is a valid name or a synonym vetted by that-and-that person in the gold standard core of the Catalogue of Life; or no, your name lives in the outer cloud, it belongs in this corner and it's probably a synonym of that-and-that species but it hasn't been checked yet. But I *do* get an answer and I know ... where it most likely belongs. And that's where we go wrong. So there is, in total, about, we have I think, 4 million names in the [Catalogue of Life]

> and there must be about 20 million names floating around in GNA and I would like to map them against a central concept, even if it is vaguely. And if there are a lot of questions coming in from the outside that say, I would like to know about the species, and they are in the outer cloud, I can actually direct my GSDs to process this part and bring it in. And the idea is to slowly consume that cloud and bring it into the gold standard so everything has been mapped. So the Catalogue of Life Plus has to do that. It has to be complete [with] all the species, about two million, and to have on board all the names that are floating around in literature (2016b).

To reiterate once again, quality taxonomies and checklists depend upon the linkages of names to documents (Taxon Name Usages) and names/concepts to names/concepts (historical variances, synonyms, etc.). All amid a word-hoard that is steadily growing and iteratively updated. The sheer size of the task, and the human-power necessary to disambiguate nomenclature ambiguities, is beyond the reach of any one organization—coordinated approaches are absolutely necessary. Organizations such as the Catalogue of Life have taken it upon themselves to collect all possible tokens and to provide a mechanism whereby user feedback can refine the nomenclatural system. What Peter Schalk's quote above shows is that exhaustive documentation and differentiation of names—while a theoretical endpoint and an ultimate goal—is not practical given both the current circumstances and the relative infancy of these projects.[63] Schalk, and by extension the Catalogue, is advocating for a multi-tiered approach, where carefully curated namespaces can exist in tandem with the created capacity to identify the range of variables in any instance of use.

In order to manage this contingency, a workflow needs to be established to stagger biodiversity knowledge according to the extent it has been curated and examined by taxonomic professionals. If we think of Bourgoin's Global Names-Catalogue of Life Parameters figure once again, we need to know exactly what name tokens are in each stage of the differentiation process at any given point in time. The Catalogue also needs to ensure that each of these 'holding spaces' for vocabularies are distinct enough that name transfers between them are carefully controlled

[63] It should also be noted that computational biodiversity methods is also a relatively recent occurrence given much nomenclatural practices trace their origins with the work of Linnaeus in the 18th century.

and documented. As a GBIF representative at the NAMES meeting conveyed, "[what I want is a] mechanism [whereby] I can give a name to some system (a name string, including those that come with a genus and barcode id) [where] I can get an appropriate taxa for it [and] tie to other data sources … [and] be given back a [vetted] token… The names are nothing to me in their own right. Names to workable concept[s]. Names linked to taxa are what are important." (personal communication, November 10, 2016). The Catalogue of Life Plus Layer Schematic (Figure 13) is a draft illustration of what Dr. Schlock describes above, and one that provides the workflow mechanisms necessary to manage the nomenclature process as described by this GBIF representative.

The outer level of the Catalogue of Life Plus Layer Schematic is the pool of Operational Taxonomic Units (OTU) designations and various other genetic sequences that haven't yet been readily reviewed and matched to traditional Linnaean species concept name designations. An OTU is a computationally taxonomic unit—somewhat equivalent to a species name in the Linnaean sense in that the token represents a conceptual taxon unit—quite common in the examination of microbial biodiversity (He et al., 2015), but also used within other biological domains to designate species sets within a group of genetically tested individuals. An OTU name, or designator, represents a "taxa yielded by grouping of specimens through a set of [genetic] markers" (Blaxter, 2004, p. 2). The issue here is that an OTU name (represented by a genetic code token, for example) cannot easily be corresponded to species concepts as they are expressed in Linnaean nomenclatural forms (in traditional scientific names) or the taxonomies these names are embedded within. For the Catalogue of Life Plus to be able to match these OTUs with their traditional scientific name counterparts, the system will require a kind of informational "hook" attached to it, such as a "code, sequence, specimen, or mixture of specimens" that can be

connected into the already-existing knowledge network. This activity will take a great deal of intervention, but is seen as quite essential if the Catalogue of Life is going to be able to move forward as *the* nomenclatural core of the iLife ecology. As one NAMES meeting attendee indicated, "In the past we have used a paper based approach to catalogue life. If we want to accelerate that and carry out this work forward, we need to give [taxonomists] a robust framework to deal with concepts. Some of this stuff around the edges, particularly sequence-based, is essential." And even though scientists are decades away from applying names to this (growing) set of data, building a mechanism for solving the widening gap is key. If true *exploitative* power is going to be gained from the system of documentation these issues needs to be rectified.[64]

Another group of textual objects mentioned by participants that can live in this outer "OTU" space of the Catalogue Plus are vernacular and common names. Up until now, while the "common name" field is "obligatory, if available" (Species 2000, 2014) in the Catalogue standard dataset (see Figure 8, chapter 1), questions remain as to how much effort should be focused on *the exhaustive completion* of vernacular names as a priority for the Catalogue (or any global nomenclator for that matter). Some meeting participants indicated that the system "should accommodate them when important, but not prioritize it general," while others felt the task was far too complicated to take up as part of this initial planning effort. One scientist noted that informal, vernacular names shouldn't be treated as just another OTU identifier, especially given the fact that any formally named species would have multiple common names in numerous languages. Additionally, there would be no way to assess how many names *potentially* constitute this vernacular space, so thus no proper goals, or assessment of progress, could be articulated.

[64] See chapter five for more discussion of taxonomic systems that fall outside the boundaries of the Catalogue of Life.

Even as scientists disagree on the number of species currently on the planet, at least some mechanisms are in place to estimate these numbers. It was generally decided that vernacular names should be left as-is and connected to the validated data when/if the occasion calls for such linkage *at the point of use*. The point of having an outer pool of text strings is so that *potential* linkages can be made in the future without having to set in place expert time and effort in less prioritized data points.

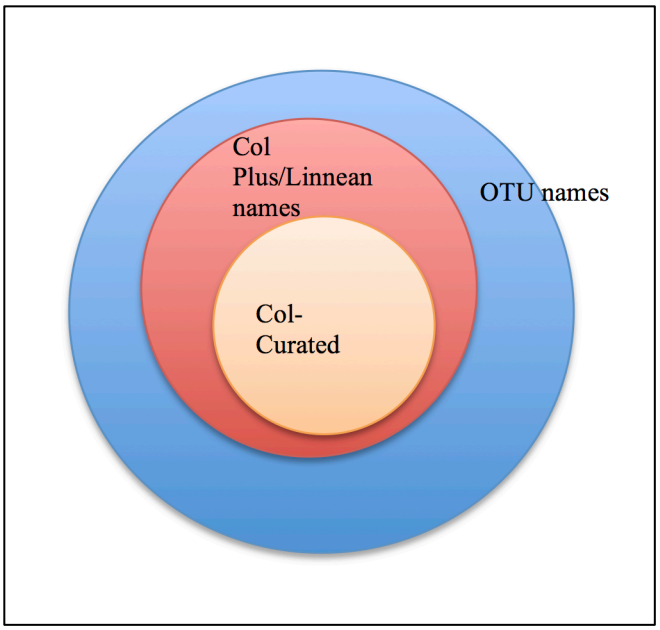


Figure 13. Catalogue of Life Plus Layer Schematic

Moving progressively inward in the Catalogue of Life Plus Layer Schematic, the middle section is "Col Plus/Linnaean names" section, which contains Linnaean/scientific names that have not yet been brought into the formally curated center section. This area "represents all scientific names (Linnaean names) that would be covered in a nomenclator and are either available or unavailable and also all subsequent name combinations. The reference to Linnaean names, [is] to separate it from names that aren't Linnaean (i.e. OTU names)" (T. Orrell, personal communication Feb 3, 2017). Though much of this landscape was already covered in the

previous section of this chapter, this is the nomenclature infrastructure component that overlaps with Bourgoin's Sections 2 and 3 in Figure 12. One essential difference here is that names in this middle layer of Catalogue Plus may not already be contained within GSDs that are currently contributed into the Catalogue. Recall that the Catalogue is built by the compilation of over 160 global databases—all of which already have a certain level of nomenclatural and taxonomic curation. There are species and whole taxon groups, however, that have no GSD actively examining their status—perhaps a GSD has not yet been created for that particular taxon (less charismatic species get less attention), or perhaps a just discovered species hasn't yet received any research coverage. This middle-layer provides an opportunity for the Catalogue to collect scientific names that would not otherwise be collected through currently existing GSD sources, certainly important as they strive to address gaps in taxon coverage and geographic regions.

Finally, the inner-most circle of Figure 13 is the Catalogue of Life in its current form and "represents scientific names in standing (available or valid based on code) that have been organized taxonomically" (T. Orrell, personal communication Feb 3, 2017)—what Peter Schalk referred to as the "gold standard core" of the Catalogue's structure. The end result of this full schematic is that names could progressively move from the outer layer to the inner-layer over a period time, taking advantage of *both* the Catalogue of Life's editorial and scientific staff as well as the users of the system, many of whom are experts in their particular areas of study.

Conclusion: Fixing Complex Concept Objects

The process of nomenclatural control, then, is the process of merging the power to *describe* documents for adequate collocation with the power to *exploit* those same documents for one's own scientific knowledge-generating purpose. And while names may not be the "perfect instruments for indexing" (David J. Patterson et al., 2006, p. 370), the Catalogue is certainly

making a good deal of headway by structurally networking *complex concepts* to the shifting formulations of nomenclature. The documents that represent species concepts within databased infrastructure, however, will always be evolving and contingent, and the mechanisms scientists create need to be equally as flexible to account for this continuing shifting. As it stands, and as we saw in the previous chapter, intermittent database documents *must* be fixed to make them function as citable, circulateable entities. An inherent tension arises from these two documentary states.

In Bernd Frohmann's "Revisiting 'what is a document?'" (2009), he points to Linda Zirelli's assertion—channeling Wittgenstein— that the concept of a document need not be *fixed* in order to function effectively and properly as units of communication. Frohman states, "our concepts can lack fixity…what counts as following a rule can be multiple, yet we can still communicate and speak meaningfully" (2009, p. 294). The context in which *specificity* functions here is important, especially given such *production* of meanings are "as we speak" (2009, p. 294). *Systems*—taxonomic or otherwise, but certainly computational—cannot interpolate minute divergences of meaning as part of the process of communication and retrieval. The tripartite interplay between names-publications-and-types must be fixed in the Catalogue in order to facilitate ready use. *Relevance*, as we well know, must be *constructed* by some logical formulation of recall and precision (algorithmic, descriptive, etc.). Fixity, rules, and, in specific, a well-bounded *species concept* is absolutely essential in the biodiversity computational field. But such fixity is in contradistinction to the *fluidity* of names, as Remsen makes clear,

> Identifiers such as names have utility in information discovery and retrieval that is directly proportional to the degree of correlation between the term and the associated meaning or, in the semiotic context, in the correlation between syntax and semantics. Laypersons may think of scientific names as stable and unique, where a single Latin binomial name refers to one species and remains that way for all time. In other words, that there is a stable one-to-one relationship between a name (syntax) and the taxon (semantics) that it labels. This is an important informatics pre-condition if we are to rely on names as a means to search for and retrieve relevant information related to taxa (2016, p. 211).

One of the goals of the Catalogue is to impose a sense of order within a nomenclatural landscape that is defined by conceptual fluctuation. Names (and thus the species concepts these names are associated with) are preserved and *fixed* within an ontologically distinct system. Such knowledge facilitation, however, even with the best nomenclator mechanisms in place, distances a *name* from the *complex,* organic *concept* as it exists at the conceptual level, making it incredibly difficult to reconstruct what Remsen called the "congruence" (2016) between a name and taxon concept. Such distance also precludes us from being fully able to assess the veracity of any one system as it applies to a particular purpose at the point of use. Recall Timothy's Utteridge's ultimate statement about databases: *you still need to go to the source*. The Catalogue (and any database) necessarily flattens and rearranges names in order to share more adequately and seamlessly. A core issue in understanding how this flattening is taking place is to acknowledge that biodiversity taxonomic systems do not deal with just one *document* when we speak of a species, but rather must conflate multiple entities that together constitute a nomenclatural *complex concept object*.

The problems with the delivery of information at scales as large as the Catalogue is *not only* a technical hurdle to be solved by biodiversity informaticians, but it is also a theoretical hurdle that must be overcome by taxonomic experts. Theory circulates socially and is translated into practices, and biodiversity taxonomists have to reckon with the Catalogue's unique epistemological position as a data management tool more than a space of taxonomic hypothesization. Within the culture of biodiversity science, taxonomies are understood to be a theoretical position about relationships, phylogenies, and ecologies. The Catalogue upsets this view that taxonomies must be internally consistent, for it is a taxonomic mélange concerned more with data collection toward the goal of easy *access* to the sum of global knowledge.

Thus, we must now discuss the Catalogue's *taxonomic identity*. The next chapter addresses the *taxonomic instrument* in which the Catalogue's name-tokens are embedded, borrowing from Patrick Wilson's notion of the *bibliographical instrument* from *Two Kinds of Power* (1968). The Catalogue's consensus-based management hierarchy provides the structure by which names become exploitable, but it also articulates a particular *position* about how biodiversity knowledge can and should be shaped and interpreted for effective information management. However, unlike the nomenclature discussed in this chapter, taxonomy has no unified *code* to guide its production; they are subjective productions that assert positions. Taxonomies inhabit a hybrid space by being both information retrieval structures *and* intellectual (hypothetical) scientific formulations. By including a species concept into a taxonomic framework we force that *concept* into the *form of a unified document* that presents a particular point of view. Is such exploitation favorable over descriptive-oriented taxonomic modes? How can we articulate the differences between the two structures? What are the implications of this schism on how we understand the context of taxonomic document-as-instrument?

Chapter 4: Documentary Instruments: Taxonomic Specifications, Consensus, and Interpretive Flexibility

Introduction

Up until this point I have spoken primarily of "units" of information—the Catalogue-as-document, the functional documentary entities and productions of the Catalogue, as well as the documents that are used to collectively create the species concept represented as a name-token within the database. Specimens and descriptive literature provide the primary evidence for the species concepts, while names anchor that information within a system of well-formed and accepted scientific tokens. Knowing now that the documentary units of biodiversity databases can be understood to inhabit a scale from data to documents to nomenclature (as a document representation) to taxonomy, I would like to switch our focus to what it means to call a taxonomy a document, and how we can begin to frame its structure as a kind of knowledge production that is central to the delivery *and* interpretation of biodiversity information. To do this I want to expand upon Patrick Wilson's concept of the *instrument* of bibliography articulated in *Two Kinds of Power: An Essay on Bibliographical Control*:

> I cannot tell how much bibliographical control I have or could have simply by introspection, by memory of past success and failures, or by flexing my muscles. To discover what I can or might do if I would, I must discover what arrangements there are of which I can take advantage, what bibliographical instruments…are at my disposal. Which objects are bibliographical instruments? … Any text that refers in any way to any other text or copy of some text might be considered a potential bibliographical instrument; the set of texts referring in some way to other texts and copies might be identified with the entire potential bibliographical apparatus, which would then be identified with the entire potential bibliographical apparatus, which would then be a very sizeable portion of the bibliographical universe…
>
> …the essential characteristic of a bibliographical instrument is that it consists entirely or primarily of descriptive works, texts, and copies … [The] formal apparatus consists of bibliographies, lists of abstracts, catalogs (published and unpublished) of collections of writings, inventories and calendars of manuscripts, book review journals, guides to literatures and to repositories of copies, indexes to periodicals, review articles, and (now or in the future) bibliographical machines or the physical apparatus … that makes machines into bibliography producing machine (1968, pp. 55–57).

In our *documentary* universe, instruments take the shape of classifications, into which names can be organized and classified in order to be functionally accessible. In no uncertain terms,

taxonomies form the basis for biodiversity discovery; they are the primary mechanisms by which we collocate evidentiary documents, frame those documents into an interpolated (intellectual) context, and then subsequently act as the access structure for information discovery and the production of new knowledge. The species lists, and the taxonomies that they are embedded in, are essential because, without them, Wilson's powers of description and exploitation would have no apparatus in which to function and circulate. As a scientific instrument, the Catalogue serves as an "apparatus for registering, measuring, or recording a physical quantity, property, or phenomenon" ("instrument, n.," 2017), of biodiversity knowledge collocated from numerous sub-instruments (GSDs and RSDs) throughout the globe. The previous chapter displayed the rather detailed intellectual, professional, and organizational apparatus necessary to control the creation, proliferation, and use of nomenclature.

But, as Wilson makes clear, in order to make full use of these instruments, we must understand how they work and the reasoning of their complex internal structures and compositions. The ordering and imposition of structure onto names is, as we have seen, a controlled, iterative process. Rules regulate the ways in which names connect to, and are formed, by concepts—based on literature and types—and while any reasonable set of biologists may disagree as to the interpretation of these assessed documents, clear rules are in place by which names/concept can be reinterpreted, assessed, and validated. Far more problematic are the final steps in Dr. Bourgoin's schematic (again, See Figure 12), which are a) the structuring of names into groupings of valid taxa (the act of classification), b) the imposition of a management classification (the Catalogue of Life itself), and c) the use of the management classification data to hypothesize and understand and postulate broad evolutionary and phylogenetic trends. These final steps mark the clear demarcation between organizing names into circulateable species

concept units, and how these units get subjectively aligned and composed within *instruments* of taxonomic *knowledge* and expertise. Steps a) and b) are the subjects of this chapter, while the latter step will be reserved for the next.

Some pertinent questions here are: In what forms do classificatory schemes take in different taxonomic subdomains and what *kinds* of emergent information do they communicate as new forms of documentation? What are the resultant structures by which this information can be browsed and accesses? How does the Catalogue's structure, as a system primarily defined for information retrieval and access, differ from traditional taxonomies that document and describe internally-consistent taxonomic positions? And finally, how does the Catalogue's management classification, intended to be reused, shared, and manipulated into various new contexts, *extend* Wilson's understanding of bibliographical instruments and the power they provide a user of information?

Taxonomic practice is defined by competing approaches, methodologies, and theoretical views; such divergence of opinion facilitates the growth of scientific knowledge over time. Given the multitude of methods by which similarities and differences are assessed between species concepts, and the extent to which (stringent, dedicated) individual judgment plays a role in any taxonomic schematic, it means that consensus on *one unified approach* is, if not impossible, incredibly unlikely. As A. Broadfield aptly stated,

> It is said that the classification of the sciences has 'proved to be a peculiar baffling problem of mountainous magnitude and of many point of view. Many thinkers had supposed from their several viewpoints that they had surmounted the problem; yet none had quite succeeded according to the consensus of scientists and philosophers regarding the order of nature.' *A consensus of scientists and philosophers regarding the order of nature would be a miracle in itself, especially if achieved without presupposing classification, which is a principle method of science and philosophy* [emphasis added] (1946, p. 70).

Due to this methodological and interpretive fragmentation, the biodiversity world is in need of a system to bring these disparate sources of knowledge together. It is these distributed taxonomic

instruments scientists use to organize (and ultimately access) names, often built at a local level, for local purposes, which we will now examine more deeply. The global success of the biodiversity discipline depends on being able to examine and assess knowledge collectively. As taxonomic expertise dwindles (Hopkins & Freckleton, 2002), and as positions supporting 'traditional' taxonomy are reduced within institutions, it is all-the-more important that data is aggregated in order to maximize its functionality and potential. [65] Taxonomies are produced everywhere, and this segmentation makes it difficult to assess the *state* of the discipline—how much progress has been made toward documenting biodiversity and what taxa need more attention given historical and contemporary scholarship.

In order to do this, we will again invoke Wilson's framework for the specifications of bibliographic instruments. Wilson identifies five *specifications* of bibliographical instruments that aptly apply to our documentary notions as well:

1. The domain of the instrument, the set of items from which the contents of the work, the items actually listed, are drawn;
2. The set of all items considered for addition to the library collections;
3. How it is determined what is to count as a unit for listing and description;
4. What information can we expect to find out about an item, given it will be represented as a unit, and finally;
5. We must understand the frequently extraordinarily complex system of arrangement or organization: we must know where items of a given sort will be found, and what it means to find an item at a given place (1968, pp. 59–62).

This dissertation has thus far examined the first four elements to a great or lesser extent–we looked at what items constituted the Catalogue's work, what kinds of documents the Catalogue

[65] My discussions with numerous biodiversity professionals made it clear that, while there is an increase in molecular and genetic approaches to taxonomic and phylogenetic organization, taxonomists trained in the careful description and naming of species are becoming less common. Chapter five of this manuscript will touch upon some of the issues that arise from this disjoint—namely, that the production of computationally- and numerically-based taxonomic knowledge is becoming divorced from the nomenclature-dependent taxonomy that comprised most of the past practices of biodiversity taxonomic and descriptive work.

collected, and what kinds of evidence these documents pointed to.[66] It is the fifth specification that we will now strive to unpack. To disentangle these complex systems, I have broken up this chapter into four parts. First, a brief section will bridge our discussion of nomenclature in the previous chapter with the concept of taxonomic instrumentation. This transition is defined by the increasing necessity of professional scientific *judgment* as part of the collocation process. This shift marks our entrance into what Bourgoin calls the domain of "taxonomic *knowledge*" (my emphasis).

Part II will center on the traditional role of biodiversity taxonomies in the sciences, and examines what *kinds* of knowledge they represent as coherent systems that express a set of assumed and implemented epistemological and classificatory commitments. Using Jonathan Furner's notion of "identity" in knowledge organization (2009), I detail descriptive-oriented and retrieval-oriented systems. While *all* classificatory systems contain elements of both, I contend that traditional biodiversity taxonomies can be characterized as *descriptive-oriented* systems, encompassing as they do the motivations to *describe* biological relationships in idiosyncratic ways and take a particular (hypothesis-driven) position about how organisms are related. These relationships represent an epistemological position about what constitutes the networked, representational (interpreted) 'reality' of a given organismic taxon or ecology. I then describe a few basic taxonomic approaches (or classificatory constructs) prevalent in taxonomy: evolutionary taxonomy, cladistics, and phenetics. This examination will illustrate how vastly different these taxonomic approaches are in practice, and how these diverse theoretical and methodological positions manifest themselves in contrasting graphical representations that can be interpreted in multiple ways. The result of this situation is that these distinct taxonomic

[66] Incidentally, these steps overlap quite readily with Bourgoin's schematic of incrementally more well-formed and truthful and scientifically available information.

representations are fundamentally incommensurable in consensus structures as-is, and require modification to collocate with other sources of taxonomic knowledge.

Acknowledging that all GSD and RSDs ingested into the Catalogue of Life represent a similar set of instrumental assumptions, Part III then switches focus to *access-oriented* systems, illustrated by structures such as the Catalogue of Life. These systems are motivated not necessarily (or rather, only) by a coherent argument about biological relationships, but rather by the need to *communicate* information among many taxonomic traditions. The result of this approach is a *composite taxonomic instrument*: a taxonomy that merges multiple taxonomic schemas into one coherent structure. In specific, I will examine the Catalogue's management hierarchy, which in biodiversity terms, is a rather drastic departure from the traditional descriptive-oriented schematic. The management hierarchy is driven by an encyclopedic ethos: it strives to chronicle the state and breadth of extent knowledge about global taxa.[67]

The concluding section of this chapter proposes to expand upon Wilson's two powers associated with *exploitative* and *descriptive control*. I argue that there is a third power that is exemplified by the Catalogue: the power of *extensive flexibility*. The terms the Catalogue has set deviate from many of the expected norms of classificatory practice, in that they are designed to be downloaded, integrated, and manipulated, into any number of online infrastructures (within

[67] In assessing bibliographic instruments, Wilson notes that the "the organizational component of the Specifications [consist] of the implicit or explicit specification of a number of *available positions*" (Wilson, 1968, pp. 62–63), into which things are placed according to pre-determined, articulated sorting mechanisms. *Things* are meant to refer not just to "physical objects, taking up space and time, but *anything*: objects, events, qualities, relations, real or imaginary" (1968, p. 65). These things also include taxonomies, which themselves represent a complex arrangement of relationships, qualities, and negotiations. Wilson provides two general *kinds* of lists in which documents can be organized: the first according to descriptions, assuming that a particular *subject* term matched that document's general content; the second is focused on utility, which "requires an estimation of what the writing is good for" (1968, pp. 66–67). Wilson equates the former kind of knowledge organization to that of a (library) catalogue, while the latter he equates to an encyclopedia. The Catalogue, built with the purpose of providing a collection of taxonomies (in the guise of GSDs) based on expert opinion, designed to provide access to *many* taxonomic interpretive approaches, can be seeing as falling into the latter category.

the iLife ecology and beyond). This *intended* ability to be implemented and repurposed in a variety of networked environments is a powerful and unique aspect of biodiversity management classifications. In Wilson's universe the *instrument* was not, necessarily, seen as a repurposable document, but in the networked ecology of the biodiversity world, the instrument itself becomes a circulating structure. This is a very powerful aspect of the Catalogue, and a space that situates it to influence the shape of biodiversity knowledge in radiant and deeply rooted ways. The *control* of the taxonomic space in the Catalogue is less about internal consistency and more about the ability to forfeit this consistency in an attempt to maximize its *en bloc* usefulness in other systems. Paradoxically, the relaxation of control also increases the exploitative power to the entire documentary universe.

In *Two Kinds of Power*, Wilson states,

> "There is a distinction between not finding what we are looking for, and finding what we are looking for is not there; the former is a failure, the latter a negative success. I do not discover the full extent of my power by reflecting on my positive successes, my occasional finding of what I seek; I must be able to recognize negative success as well as distinguish them from failures to do what might have been done. I cannot make the distinction accurately, however, without knowledge of the Specifications of the instrument, the rules according to which it was constructed. (1968, p. 59)

With this in mind, let us now examine the biodiversity taxonomic instrument in some detail with the ultimate hope that it can help us better understand their specifications. Such an understanding can help us see the potential power of consensus-based systems as vehicles for more integrated—and, ultimately, pluralistic—knowledge organization systems.

Part I: From Nomenclators to Instruments of Knowledge

Before speaking of taxonomies in particular, I would like to bridge our discussion from the last chapter—that of nomenclatural control based on objective rules of priority—with the stages in Thierry Bourgoin's model (Figure 12) that constitute more knowledge-based activities requiring subjective assertions about taxon. Moving along the Global Names-Catalogue of Life

Parameters schematic, the point at which information control begins to make this transition is between stages 3 and 4, when the creation of "valid" taxonomic knowledge begin to take shape in the form of *classifications*. Up until stage 3 the focus was on nomenclature, and as stated by an interview participant, "Nomenclature is not really interesting, it's full of rules. It's like [the law]—[lawyers] defer to a set of rules. [They] get [a] book out and in 1925 so and so is precedent for this." The core sentiment of this statement is that names, built as they are for stability and consistency of use, are created and accepted based on very clear guidelines: a name is either published or it is not, a type is either described in relation to that name or it is not, one name was either published before another or it was not, etc. Nomenclaturalists do not make judgments; they follow a set of rules. As the ICZN makes clear, "The Code refrains from infringing upon taxonomic judgment, which must not be made subject to regulation or restraint" (International Commission on Zoological Nomenclature, 1999). This distinction is crucial to understand, for it is within the spaces created between nomenclature and taxonomic judgments where biodiversity instruments take their shape.

As part of this transition from nomenclature to taxonomy, names are concatenated into groups that, after reconciling potential species concepts, assure *one instance* of a species concept per scientific name based on the rules from the applicable code of nomenclature as well as the application of evidence to decide upon the boundaries of taxa. Deciding that one species concept is correct over any other, however, is a matter of expertise. As has been illustrated, different names can be applied to taxa depending on the taxonomist that sets its boundaries (International Commission on Zoological Nomenclature, 2016b). Recall Richard Pyle's example in the last chapter illustrating the complex development of what began as the *Holocanthus fisheri* (Barnthouse 1904) species. Over time, different scientists provided different specific boundaries

around taxa, and each of these interpretive acts resulted in the production of an equally-valid and available name and species concept. The nomenclators that collect these various nomenclatural acts are simply "definitive listings of code-governed names, their orthography, and bibliographic citations" (D. J. Patterson et al., 2010, p. 3), and nothing more. Taxonomy, however, is a distinct and separate process altogether. Simply put, taxonomy is a *science*. While nomenclature follows clearly articulated procedures to produce a set of circulating facts, taxonomy orients these facts in different argumentatively meaningful ways.

As Bourgoin's schematic helps us understand, the informed *judgment* of a scientist becomes more important as names and taxa begin to get mapped and related. Not all token relationships, however, are equally easy to assess. Taxonomists can negotiate the transition between nomenclature and taxonomy well enough because they have access to documentary evidence and are trained in their craft—this, after all, is what defines their work. Biodiversity informaticians, however, "are encountering unfamiliar problems that confound the merger of distributed data" (D. J. Patterson et al., 2010) since these taxonomic reconciliations cannot be easily mediated and solved computationally.[68] Managing these variances within a biodiversity database is a difficult task. Both the "many-names-for-one-species" and one-name for-more-than-one-taxa problems (D. J. Patterson et al., 2010, pp. 3–5) test the limits of what databases can handle given the vast amount of data being coordinated in database infrastructures such as the Catalogue. As a general rule, what falls outside of the domain of a nomenclator are "subjective assertions of heterotypic synonyms and higher (above genus) classifications –

[68] We will see an example in chapter five of how GBIF is creating taxonomic hierarchies with the use of algorithms.

basically the arrangement of taxa in a classification" (T. Orrell, personal communication, February 6, 2017).[69]

The difference between homotypic synonyms (or objective synonym) and heterotypic synonyms (subjective synonyms) can help illustrate this general issue. Homotypic synonyms, as was briefly mentioned in the last chapter, are different names that refer to the same type of material(s) at the same taxonomic rank. So long as unambiguous type information is clearly identified in contributing GSD fields, creating a database relationship between the name-token entities that represent them is relatively straightforward. Within the ICZN, homotypic synonyms are appropriately called "objective" synonyms, primarily because their identification is contingent upon the objective use of the same type specimen; there is no special expertise necessary to make this connection within the database.

Heterotypic synonyms are another matter entirely. Heterotypic synonyms within the ICZN (or "taxonomic synonyms" in the botanical code) require more specific and expert judgments. Heterotypic synonyms occur when a scientist infers that two "code-compliant scientific names … are thought to refer to the same species" (Global Names Architecture, 2017d). Heterotypic synonyms are referred to as "subjective synonyms" (International Commission on Zoological Nomenclature, 2016a) because they are based on the independent judgment of a scientist who has examined both the type specimen documentation, as well as the published circumscriptions that denote the identification of that species concept; and concluded that two species are, in fact, the same (or different, at which point a new name and circumscription must be applied to one of the taxon groups). It is this balance between nomenclatural control and taxonomic opinion that allows taxonomic work its *flexibility*

[69] See also (Blake, 2011, pp. 468–469).

(exhibited by Pyle's fish example in the previous chapter). Controlled concept vocabularies, type specimens, and publications, provide the concept-tokens that can be subsequently re-combined in flexible ways as taxonomists delimit concepts within specific arrangements (see Figure 14).

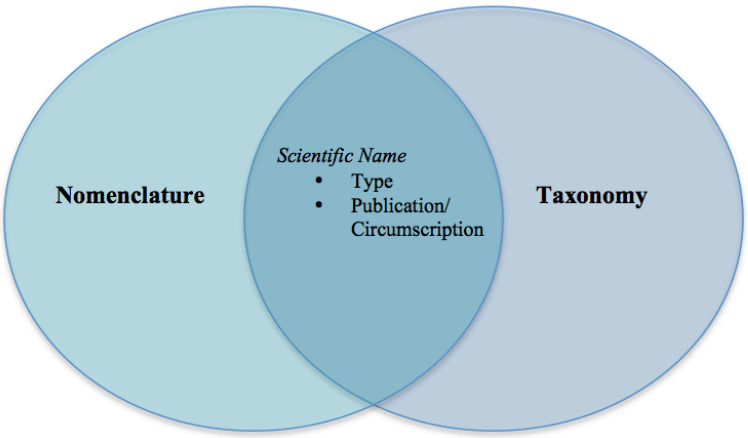


Figure 14. Overlap between the task of nomenclature and of taxonomy, with the type and circumscription being used for both assessments, though in vastly different ways.

The result of this synonymic differentiation is the production of a hierarchy of terms that differentiates an accepted name from its synonyms and other associated nomenclatural data (See Figure 15, below).[70] Name hierarchies provide token groups that delineate *valid taxa* that can then be combined and situated alongside other taxa to create a taxonomic classification for the organization of life. For example, *Sabella discifera* Grube, 1874 in Figure 15 is a *curated* unit of scientific taxonomic knowledge that is surrounded by a number of synonyms and sources that, together, constitute an interpretation of how this taxon concept was intellectually constructed by a taxonomist. The World Register of Marine Species (WoRMS) (2017b), a self-described "authoritative and comprehensive list of names of marine organisms," is careful to note that

[70] A number of other nomenclatural issues arise as part of the taxonomic process—which can be seen in Figure 15 with the identification of "invalid combination" and "original name"—but I have chosen to focus on only a couple of examples to illustrate my general point.

"taxonomic experts" and not "database managers" are involved in the editorial management of the platform. Such a designation is important since curatorial intervention is seen as far more precise and accurate process than computational methods.[71]

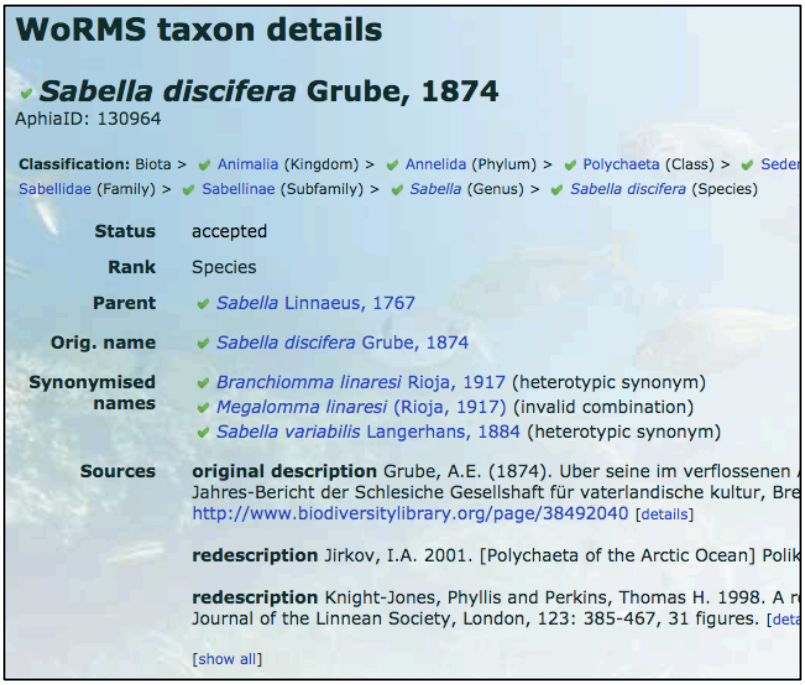


Figure 15. Nomenclature hierarchy for the sedentary Annelid (phylum consisting of segmented or ringed worms), *Sabella discifera* Grube, 1874, provided by the World Register of Marine Species database. Note the synonymized heterotypic names, each of which required expert examination to determine that these were, in fact, the same species (World Register of Marine Species, 2017a).

Referring back to Floridi's sense of truthful information, while names have codified mechanisms by which they can be accepted as "known-to-science" factual statements, once concepts start getting paired to them, opinion and hypothesizing take over as the dominating paradigm. Taxonomies are not *facts* in the sense that names represent validly circulating concepts—they are hypotheses (Thiele & Yeates, 2002), and mechanized, well-crafted suppositions. A species concept, as well as the relationship between them, are always bound to change as new interpretations are postulated, as new methods of assessment are introduced, as

[71] WoRMS is highly-respected in the field of biodiversity taxonomy, precisely because it is an RSD that spends a great deal of time curating its content with support from a number of experts from around the world.

new species are discovered and described, etc. The challenge then becomes how to construct a biodiversity taxonomic database that manages to mimic the flexibility of *practice* within the technical infrastructures themselves. Taxa *bring* knowledge to database domains as distinct packets of historical and expert negotiations; taxonomic databases like the Catalogue of Life *structure* that information in ways that make it easy for individuals (and systems) to navigate and share that information on a wider level.

Some distinctions need to be made here before we proceed, however, since I have woven together a discussion of 'traditional' taxonomies with that of the Catalogue above. There are *two* kinds of taxonomies that will be discussed here on quite separate terms: first there are those taxonomies that do, in fact, represent particular unified, consistent phylogenic and taxonomic opinion. These taxonomies exemplify a *descriptive-oriented* paradigm. Within this project, descriptive-oriented taxonomies are generally represented by the global species databases (GSDs) that come together to form the Catalogue's taxa content. These coherent structures can, theoretically, be represented by any number of classificatory schools, including evolutionary taxonomy, pheneticism, cladism, etc. These different approaches will be ever-so-briefly described below to illustrate how their individual productions conflict with each other, both in theory, as well as in their resultant representational structures. In other words, in addition to a documentary universe, there is also a universe of all taxonomic interpretations and documentary instruments. Instrument specifications are also *contingent*, much like the documents that they contain. Second, there are classification systems such as the Catalogue of Life that attempt to reconcile these *numerous* taxonomic opinions into *one* coherent structure as a means of data access. These taxonomies exemplify a *retrieval-oriented* paradigm.

This chapter will now focus on these two types of classifications to distinguish what is *unique* about the Catalogue's status in the biodiversity community. But first, a brief sojourn to help situate what the primary functions of taxonomies are in practice, and the essential differences between descriptive-oriented and retrieval-oriented approaches to classification. To do so, I will invoke Jonathan Furner's notion of descriptive-oriented and retrieval oriented approaches to the goal of knowledge organization systems (2009).

Part II: What is the Function of Classification?

Entire monographs and numerous textbooks in Information Studies have and will continue to be dedicated to what kind of an instrument a taxonomy *is*, how they are built, and how they can be maintained. Classification, as Henry Evelyn Bliss indicated, is an *action* (to class… "Assigning a thing or several things to their respective classes"); an *act* (to *classify*, as in to *conceive* of "classes in some order and relate them in some system"); and a *product* ("a series of system of classes arranged in some order according to some principles or conception, purpose or interest, or some combination of such") (1929, pp. 142–144). My focus in this section is merely to provide the briefest entrée into this arena of discourse, by thinking about what the *functions* of a taxonomic instrument are within a domain like biodiversity studies using these general concepts. The basic premise for this distinction lies in the fact that each approach—descriptive- or retrieval-oriented—to *building* a distinct taxonomy requires a different epistemological understanding of how a taxonomic classification is supposed to function in biodiversity spaces, and how it is that we can judge whether or not they succeed at what they are designed to accomplish. It is important to emphasize that there is not one unified approach to producing a biological classification; a classification *is* the scientific argument, and therefore

each production is a unique structure, layered with assumptions at every minute level that we've so far discussed. Documents arranged are only documents-proposed.

Taxonomic arguments change over time. Taxonomic knowledge systems are products of contextually specific historical, cultural, and philosophical circumstances that evolve over time. Insofar as the broader purpose of KO systems is to organize any given domain of knowledge or documentary production, their aims are also "to achieve consistency—to produce identity—between (i) the KO systems designer's representation of reality, which basically amounts to the aggregate of all extensions of all subject classes and the relations between them, and (ii) the KO system user's model or image in the world" (Furner, 2009, p. 12). The end result of a KO system is inextricable from the epistemological and classificatory commitments and expectations of both those who build these systems for the organization and retrial of information, as well as those who wish to harness the "exploitative power" (Wilson, 1968) that KO systems provide. Taxonomies are artifactual in this sense, in that a close examination of the way they compose and *document* knowledge can tell us something about the "the social organization of knowledge on the one hand, and on the other hand the intellectual or cognitive organization of knowledge" (Hjørland, 2008, p. 86). But how do we assess this artifice?

Jonathan Furner proposes one way that we can potentially assess knowledge organization systems by asking how accurately they represent the external world, as well as its internal coherence and simplicity,

> I think it is possible to distinguish two conceptions of the goal of the practice of KO, and this distinction corresponds roughly to the one Raya Fidel draws between two conceptions of the goal of indexing. On the one hand, we have the document-centered view that indexers should aim to assign index terms to documents (or documents to index terms) in whichever way it is that produces the most accurate representation of that content. On the other hand, there is the user-centered view that indexers should aim to associate documents with those terms that are most likely to be used by searchers looking for those documents. Similarly, I think that, on the one hand, we have a description-oriented conception of the goal of KO, being to build systems that do well at helping people produce accurate descriptions and representations of documents. And on the other hand, we have a retrieval-oriented conception of the goal of KO, being to build systems that do well at helping people find the documents they think they want to find

(2009, p. 9).

This distinction between descriptive- and retrieval-oriented approaches to knowledge organization seems to me an apt way to think about the kinds of taxonomies that flourish within the biodiversity world. As Tony Rees, Manager of the Divisional Data Centre, CSIRO Marine and Atmospheric Research in Tasmania, articulated on the Taxacom biodiversity: "I look upon biological classifications as serving two purposes - first, to illustrate our current best guess/es as to the relationships between organisms, and second, to provide a recognisable navigation structure so that persons entering the classification can (hopefully) find their way to their particular organisms of interest" (2009). This quote brings to light the two distinct aspects of the biodiversity instrument. For Furner, these two approaches are rooted in the larger question about how one is to *evaluate* (Furner, 2009, p. 4) knowledge organization systems as systems that "represent relationships of identity between classes of documents," and "help people find the right labels for classes of documents that about those identities, and help people find those documents" (2009, p. 4). While there exists no standard by which the *true* "goodness" of a KO system can be quantitatively assessed, there are number of qualities that have been identified in order to critique the effectiveness of KO schemes. Furner indicates that the basic role of a KO system is to adequately represent the identity of some external reality: "The main claim that I would like to make about the importance of identity for KO is not that an understanding of identity is helpful in analyzing the structure of aboutness and relevance. It is that there is a sense in which identity is actually the goal of KO" (2009, p. 12). The rubric for assessing whether or not KO systems 'work well' can be distilled to a series of factors within either a "description-oriented" or "retrieval oriented" notion of KO (2009, pp. 9–10).

According to Furner, description-oriented KO asks two basic questions of a designed KO system: (1) How correct, just, and fair, is any given ontology in relation to the natural, real

world, and (2) How internally coherent is the infrastructure itself is in exemplified a unified ontological system with an internal logic (2009, p. 9). In the biodiversity realm this means creating classification that provides a consistent model that includes a fair and accurate representation of biological organisms, and that provides a classificatory system that depicts things the "the way things really are, or the way somebody thinks things are" (20019, p. 9). Retrieval-oriented KO, on the other hand, focuses on the KO infrastructure's ability to facilitate the location of documents or required information (2009, p. 10). Terms Furner associates with this level of assessment include, effectiveness, efficiency, and usefulness to the user. In the biodiversity world, this would mean the ability to locate species concepts and the associated species documents easily regardless of their descriptive position relative to the *user* of the taxonomy. The methods by which these elements are measured against any KO system is an entirely different matter, and up for debate. "Different people see reality in different ways, and draw from that the conclusion that every KO system is necessarily and unavoidably 'biased,' in the sense that every KO system reflects the view of its designers" (2009b, 9). Nonetheless, these elements proposed by Furner provide us with a useful starting point by which we can conceptualize the efficacy of KO systems as *both* accurate tools for retrieval *and* spaces for hermeneutic articulations.

The task for this section is to think about how classification systems function on two very basic levels: as a tool for *retrieval* (a product that has an intended purpose of maximizing a document's exploitative power), as well as a taxonomy as a tool for *description* (a document as a series of classificatory *actions* that explicitly represents a particular view of how organisms are related). My intention in making this bifurcation is to emphasize two basic approaches to knowledge organization in biodiversity sciences: one as the product of individualized scientific

work and *hermeneutic* development, and the other as a space for composite, unified information access and communication. This boundary tends to mark the divide between those that support or do not support a generalized taxonomic model such as the Catalogue of Life.

To be sure, Furner notes the division between the descriptive and retrieval-oriented approach is artificial, for these approaches often comingle in practice. Nonetheless, it is a useful exercise to bring to the fore distinctions between these two *kinds* of circulating taxonomies in the biodiversity world and what they are, and are not, intended to represent and offer. The basic question becomes: What are taxonomies *intended* to represent as vehicles for producing and organizing knowledge? Referring once again to Bourgoin's schema in Figure 12, it's important to note that the production of valid taxa coincides with the production and act of classification, but that classifications are first imposed within the global species databases (GSD) *before* the Catalogue of Life even begins to enter the systematic workflow. Artificially dividing up descriptive- and retrieval-oriented modes of classification (as much as these functions comingle in practice) makes sense operationally within the biodiversity arena since GSDs and the Catalogue *functionally* serve purposes that correspond to this division. Thus, I am going to first describe the *descriptive-oriented* classification instrument *outside* of the Catalogue, as exemplified by Global Species Databases. I will then return to the Catalogue specifically to see how its structure veers from "traditional" modes of taxonomic expressivity in *access-oriented* articulations.

Descriptive-oriented classification modes and inherited instrumentation.

Traditional taxonomies can be defined as systems that invoke a unified methodology to provide a consistent model of the natural world. Such instrumentation must work as a functional *system* of arrangements so that a species location in a system tells you about how it operates in

relation to *all other entities* within that system. Descriptive-oriented taxonomies are exemplified by the Global Species Databases that functionally provide the core data for the Catalogues management classification. GSDs are those databases that (typically) collect a single taxon group (usually at the genus or family level), and attempt to do so on a global scale—all instances of that taxon regardless of geographic boundaries. The Catalogue of Life defines GSDs as "[aspiring] to the following properties":

- cover one taxon worldwide
- contain a taxonomic checklist of all species within that taxon
- deal with species as taxa, and contain synonymy and taxonomic opinion
- have an explicit mechanism for seeking at least one responsible/consensus taxonomy, and applying it consistently
- cross-index significant alternative taxonomies in their synonymy (Species 2000, 2015e).

Because GSDs tend to cover *one* taxon (See also, Species 2000, 2014), and many are curated by small sets of individuals (or single individuals), they typically follow an internally-consistent taxonomic structure, and thus follow a formal set of directives as to how they should and can be structured (personal communication, November 10, 2016).[72] Crucially, "taxonomic opinion" in the space of a GSD is unified and not subject to multiple points of view.

While central databases such as the CoL collect subsidiary data structures to enlarge their taxonomic hierarchy, each of these taxa-specific taxonomies brings with them an inherited set of instrumental qualities. "The source databases are diverse in their origin, their purpose and

[72] Regional species databases (RSD), on the other hand, contain data that comes from specified regional boundaries, and the sources for these data comes from publications that document the flora of North American, the flora of China, European flora, and so on (Roskov, 2016b). ITIS is one example of one of these "mega-diverse regions" (Species 2000, 2015e), along with Species 2000 China (Species 2000 China Node, 2016), Atlas of Living Australia ("Atlas of Living Australia," 2016), New Zealand Organisms Register ("New Zealand Organisms Register," 2016), WoRMS (World Register of Marine Species, 2017b), and Species 2000 Europa (Species 2000, 2015e). GSDs are generally ontologically distinct databases that are internally cohesive in their method of representation, as well as in terms of how *relationships* are assessed between taxa (represented as name tokens). Unlike GSDs, RSDs, cover multiple taxa within one infrastructure. To this end, they are more akin to the Catalogue of Life model of taxonomic management: multiple taxonomic opinions comingle within one basic backbone infrastructure. Thus, the findings indicated below related to the CoL will apply to the RSDs as well, albeit at a more localized (geographic) level, so for the moment, the discussion will place RSDs on the backburner. RSDs will also be discussed as an important entity in bridging taxonomic gaps in the Catalogue's structure.

therefore their structure. A key challenge for the Catalogue of Life has been the integration of this disparate data, and a standard dataset has been established for that purpose" (Species 2000, 2015e). A key quality is the relationships that are created between names in the taxonomic environment, and what Wilson calls the "rules of assignment" for how certain concepts inhabit a specific set of positions within a bibliographic instrument (or documentation instrument in our case). In the previous chapter we saw how relationships were essential in disambiguating nomenclature (synonymic and otherwise) as a representation of numerous (potential) species concepts. Our focus here, however, is what Clare Beghtol calls the "theoretical constructs" (2001, p. 99), created by a network of concept relationships. Beghtol continues, "In general, relationships in bibliographic classification systems are functions of both the syntactic (i.e. structural) and the semantic (i.e. meaning) axes of the systems" (2001, p. 101). Thus, it makes sense to begin to think about how structure (hierarchies) and networks of meaning are *constructed* in the process of hypothesizing biodiversity taxonomic relationship schema.

Reality as an evolving representation: No universals in biodiversity.

First and foremost, because GSDs are "diverse in their origin," it is important to note that no two GSDs are exactly alike; diversity of opinion is a natural part of the taxonomic landscape. Recall Clare Beghtol's statement that "every classification system is a theoretical construct imposed on 'reality'" (2001, p. 99). On the one hand, taxonomies must maintain a certain degree of internal integrity with regard to how they represent the external world they intend to document. GSDs gain their credibility, after all, to the extent that their classifications maintain their classificatory commitments to their particular conception of what biodiversity structure represents, some notion of an external reality. As was identified by Furner (2009, p. 9), how we judge the success of a descriptive-oriented knowledge has been a question of great concern to

those seeking to classify knowledge, particularly in relation to how effectively they represent external world. Furner identifies a few more as possible criteria by which we can judge whether a particular organizing schema is successful: "coherence, richness, simplicity, or elegance," (2009, p. 9). GSDs are *coherent* in that the approach to constructing relationships is uniformly used within the classification; they are *simplistic* and *elegant* to the extent that their taxonomic tree arguments are parsimonious and describe "the evolution of any particular set of characters using the smallest number of evolutionary changes" (Wiley & Lieberman, Bruce S., 2011, Chapter 6).

These attributes aside, arguably, the most important quality with which we can assess classifications is the "degree of correspondence" between the model of relationships with "the way things really are" (Furner, 2009, p. 9). What are the *real* things in this case? The only "facts" that taxonomic science acknowledges in practice are those valid names that serve as the building blocks for classifications; the only "real" things are the *taxa* these names represent as delineated in the circumscriptions articulated in publicly-accessible journals. Everything else is a mere hypothesis (as well-supported, well-articulated, methodologically-controlled that hypothesis may be). And given the extent to which valid names (and their associated species concepts) are constantly changing and redefined in relation to the external species they represent (via a type specimen), a stable, functional reality is a shifting and elusive concept. As we've seen, taxonomies in the biodiversity world each bring with them their own set of epistemological approaches and idiosyncratic structure; no two taxonomies can be judged on the same merits or criteria. Reality is *composed*, individualistic, and, unique.

Henry Bliss (1929) takes a somewhat broader, socio-cultural view of what classification strives to represent. To Bliss, accurate representational systems did not strive to match the real

world (that is to say, organisms as they are *actually* connected), but rather classification in the sciences strives to represent the *production* of scientific knowledge as it corresponds to the study some external reality. This distinction is significant. Classifications in this sense are intended to represent the identity of the social organization of science as a whole, for "the order of nature" (Bliss, 1929, p. 170) was a function of the output of the processes science implements to document it. As he stated in regard to the organization of knowledge in libraries, " the more definite the concepts, the relations, and the principles of science, philosophy, and education become, the clearer and more stable the order of the sciences and the studies in relation to learning and to life; and so the scientific and educations consensus becomes more dominant and more permanent" (1933, p. 37). According to Bliss, accurate representation is the iteratively clearer articulation of the concepts science creates to understand nature. "A concept is …not the object of knowledge, but the form taken by the knowledge of it" (Broadfield, 1946, p. 75). Bliss's approach to accurate representation in knowledge organizing systems is one where truth is based on the accumulated knowledge amassed as part of the general empirical endeavor: the "relative quality of knowledge veritably correlated to reality" (Bliss, 1929, p. 129; quoted in Mai, 1999, p. 550).[73] Jens-Erik Mai expands upon this,

> Since Bliss regards both the mental sphere and the external world as organized in some way, he defines truth to the "the relative quality of knowledge veritably correlated to reality." This means that provided that more and more people have the same experience with reality, then this becomes the truth about the world. This implies that truth depends on external physical realities, and that the physical world is primordial to

[73] I should mention here that Bliss's approach to knowledge organization was not originally intended to be used in relation to the individual taxonomies produced *within* the discipline of any science. "Classification *of* the sciences is distinguished from classification *in* the sciences," Bliss states, "The system of the sciences comprises not only a classification of the sciences but many classifications within the several sciences. With these latter we are less concerned in this book" (1929, pp. 236–237). Bliss's conception was to create a *universal* system of classification whereupon all produced knowledge could be scaffolded. Nonetheless, I think this general approach can tell us a great deal about one way in which we can go about approaching the organization of biological knowledge. Bliss's approach is pragmatic; in the end what matters is how knowledge circulated in the domain of science, and how that collective endeavor got us closer to the "same understanding of reality" (1999, p. 550) as part of this process of negotiation.

> mental understandings of it. Our mental understandings of the world and the truth of the world is derived from our perceptions of it" (1999, p. 550).

One issue to note here, and something the study of the Catalogue specifically helps illustrate, is that our collective knowledge of the world does not *only* come from out perception of it. Our understanding also comes from the naming and knowledge structures we create to organize our previously recorded perceptions, for such structures ultimately influence and shape how and why we see what we do. The attributes we use to define a species change over time. As Gaston and Galison state,

> Linnaeus's ways of looking at, describing, depicting, and classifying plants were openly, even aggressively selective. Botanists must school themselves to concentrate on characters that are "constant certain and organic"; they must not allow themselves to be distracted by the irrelevant details of a plant's appearance and thereby unnecessarily multiply species (2007, p. 59).

The tools we use shape our perceptions (the "ways of looking" that Gaston and Gaston emphasize), and integrally limit the constitution of our knowledge practices and their productions. Michel Foucault expands this point, emphasizing that our descriptions of nature are perpetually limited because they are necessarily filtered through the language and discourse that render it visible to contemporary discourse (1994, p. 135). "Natural history is a science," Foucault continues, "that is, a language, but a securely based and well-constructed one: its propositional unfolding is indisputably an articulation" (1994, p. 136).

Some theorists and practitioners of classification within the study of knowledge organization have taken Bliss's notion of the "order of nature" to its (perhaps) logical extreme: organizing knowledge around the inherent order of the natural world. "The theory of integrative levels claims that the natural world is organized in a series of levels of increasing complexity: from physical particles and molecules, through biological structures, to the most sophisticated products of human thought" (ISKO Italia, 2004) (See also, Claudio Gnoli & Riccardo Ridi, 2014; Gnoli & Poli, 2004; Rick Szostak, 2008). This adherence to a static understanding of

external phenomena (consistent enough, in practice, that the organization of *all* knowledge can perpetually be conformed to this schematic), however, are of little use in domains as specific as biodiversity taxonomy where the "reality" in question is a shifting ground of nomenclature, concepts, and taxonomic arrangements. Such empirically dedicated approaches to classification presuppose a permanence and consistency to our knowledge that just does not exist, and overlooks the socially situated, culturally defined "unfolding" of our knowledge production practices. As Mai has noted, the impulse of "modern" classification systems to standardize and universalize knowledge organization systems around notions of "exclusiveness" and "exhaustivity" overlooks the application of these standards in individual domains (1999, pp. 548–551). In recent years, a more historically-informed notion of classification has taken shape, one where "unificationism" has given way to a "generation of theories, principles and methods that emphasize both the cultural and historical specificity of classification practices and their emancipatory function" (Furner, 2013a, p. 32).

Biodiversity taxonomy, and science in general, is anything but unified and universal. Given the fluid and contingent nature of the species *concepts* and entity documents of biodiversity knowledge, there should also certainly be no expectation that the instruments that organize them will be any less so. And while Bliss's general feeling was that science tended toward consensus (Bliss, 1933, p. 42), and "near equilibrium" (Beghtol, 1986, pp. 114–115), in the practical world of biodiversity taxonomy, nothing but the opposite could be truer of contemporary practice—*even with* the 'real,' natural world grounding the act of classification. Taxonomies bear the fingerprint of their producers—both theoretically (in terms of how they define a species and their relationships with other species), but also in the methodological approach they use to define those relationships. As John Dupré remarked, "the idea of science as

a project that might ultimately be completed in some grand synthesis of all natural knowledge is an understandable and perennial dream…the disunity of science is not merely an unfortunate consequence of our limited computational or other cognitive capacities, but rather reflects accurately the underlying ontological complexity of the world, the disorder of things" (1993, p. 7).

Amid this taxonomic disorder, then, one of the main goals for the Catalogue is to try to take this taxonomic disorder and comingle these instruments in some space that manages to harmonize this complexity for the purposes of *access*—embracing this complexity as an essential part of the identity of the classification itself. In characterizing the postmodern turn in classification Mai points to the understanding of classifications as being based on two general assumptions: "The first … is that there is no key denominator to understanding the world, neither in nature, truth, God, or future … the second assumption is the belief that there is nothing, ideas or thought, prior to language" (1999, pp. 551–552). This emphasis on language—echoing Foucault's notion alluded to above—the meaning of words, and the use of this nomenclature within the taxonomic community, certainly rings true for the biodiversity community. Classificatory constructions in biology are only as good as the *source* of their nomenclature. Mai's assertion that "language, therefore is not a tool to speak with, but the very social and cultural context in which the language is situated," takes on a great weight in the domains of biodiversity studies, and classifications more broadly construed. Words shift, and the taxonomic structures that relate those words in a classification need to be equally as plastic as the theories, methods, and organisms that circulate through time. Our next task, then, is to think about what kinds of expressed relationships we see as a product of descriptive-oriented taxonomic practice,

and how taxonomies construct particular biological realities via different methodological approaches.

Internal constructs and inherited interpretation.

Each global species database that makes its way into the Catalogue of Life has a particular structure that represents the *opinion* of the taxonomic experts that contribute them. Each GSD is produced under a set of social conditions. The distinctions between one condition and another represent deeply seeded theoretical and philosophical divides about how relationships can and should be built within classification systems, and how species are lumped into nested and hierarchical taxa. Below I will provide a glimpse into some of these possible conditions. The purpose here is *not* to provide an in-depth analysis of each school, or an exhaustive comparison of each, but is rather to show how taxonomic *productions* in-and-of-themselves furnish a unique set of classificatory relationships based on theoretical and empirical assumptions.[74] And these approaches are merely *possible* arrangements; over time, one approach will hold dominance over others. And more, other arrangement methodologies will be articulated as science proceeds forward and continues to build on past knowledge. How we understand the internal constructs of taxonomic instruments is also *contingent* on historical and cultural circumstances. The *classificatory outputs* of each approach represent commitments to these theoretical assumptions and create different arrangements and relationships that are fundamentally incommensurable to each other in a one-to-one relationship model.[75] In other

[74] The main distinctions drawn here are selected and summarized from Marc Ereshefsky's, *Poverty of the Linnaean Hierarchy: A Philosophical Study of Biological Taxonomy* (2007). Readers should review his chapter titled, "A Primer of Biological Taxonomy," for a more holistic view of these traditions. See also (D. L. Hull, 1988, 2001; Mishler, 2000) for some canonical and excellent broad overviews on the subject.

[75] I want to be clear here that I use the term *incommensurable* to indicate that these differing classificatory schemas produce classifications that *cannot* and *should not* be reconciled with each other. *Diversity* of taxonomic opinion is essential to the evolution and creation of more accurate taxonomic methods as new species are discovered and new

words, the *theoretical approach* to a biological classification changes the way we interpret collected data and relate together the subjects of classification, and thus, in turn, produces an idiosyncratic view of what constitutes *knowledge* in biodiversity studies. The "subjects" of classification in these taxonomies are those species concepts represented by *names* in databases.[76] The production and depiction of relationships in tree and hierarchical forms is fraught with a number of challenges, complications, and decisions (R. D. M. Page, 2012; Wiley & Lieberman, Bruce S., 2011, Chapter 4), which means that no two trees or hierarchies are alike, let alone can they be interpreted with the same set of assumptions. Acknowledging this fact is an important part of the story for a thorough discussion of the Catalogue of Life.

Marc Ereshefsky identifies four general schools of thought that have been prominent in the 20th (and now 21st) centuries: evolutionary taxonomy, pattern cladism, process cladism and pheneticism (2007, pp. 50–51). For the purposes of this discussion, I will lump pattern cladism and process cladism into one large cladist school, but the reader should know that I am glossing over very significant differences between these factions for the purpose of this analysis. I am sensitive to the minute differences between these analytic taxonomic approaches.

Broadly construed, evolutionary taxonomists believe that the emergence of new species (taxa) can occur through two distinct processes: cladogenesis and anagenesis. Cladogenesis is the splitting (branching) of a "single genealogical lineage" (2007, p. 52) through, for example, the process of occupying of new adaptive zones (a population of a species becomes geographically

---

methods are produced. In other words, the goal should never be the creation of commensurable structures, for such an endeavor erases and obfuscates the importance of diverse approaches in the production of scientific knowledge. I thank Johanna Drucker and Jonathan Furner for brining this fine, but essential, distinction to my attention.

[76] Matthew Hull (2001) notes that even the concept of species is a theory-laden concept, inherited from Darwin's notion of evolutionary development (Darwin, 1859): "the basic units in evolutionary classifications—species—must be the things that evolve as a result of the interplay between mutation and selection. Hence, our understanding of the evolutionary process enters into the formulation of classification right from the start…Evolutionary theory as a process theory determines the basic units of classification" (D. L. Hull, 2001, p. 22).

isolated from the rest of the population and adapts new genetic characteristics) (SæTher, 1979, pp. 308–309; Ereshefsky, 2007). Anagenesis, on the other hand, is the gradual change (divergence) (Ereshefsky, 2007, p. 52) over time of a species until it becomes distinguishable as a new species. The mechanisms and qualities that are used by taxonomists to assess "significant" (Ereshefsky, 2007, p. 52) enough changes to warrant a new species for anagenic change is a subjective and somewhat arbitrary process (Vaux, Trewick, & Morgan-Richards, 2016) and a source of much debate in the taxonomy arena. Such assessments on what constitutes a new species are the product of certain theoretical positions held by practicing scientists, much of which is based on whether or not they subscribe to the notions of paraphyly as part of an evolutionary schematic. Cladogenesis produces monophyletic taxa: a taxa that includes an ancestral species and *all* of its descendants (Grant, 2003). Anagenesis, however, produces paraphyletic taxon groups: a taxon that *does not* contain all of the descendants of a particular taxon (See Figure 16).

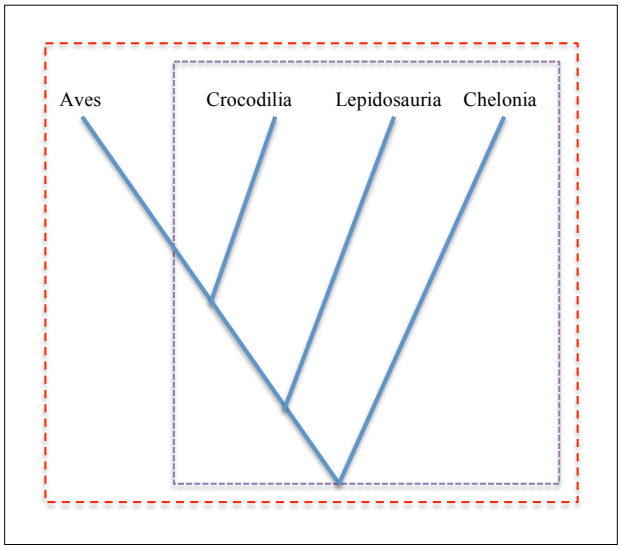


Figure 16. Example of monophyletic (outer, red box) and paraphyletic groups (inner, purple box). Evolutionary taxonomists would define the class Reptilia as only containing lizards, snakes, and crocodiles, excluding Aves as part of this schematic; thus, it is paraphyletic. Cladists, on the other hand, would be unwilling to exclude Aves because groups should include all of the descents of a taxon, and thus a Reptilia group that includes Aves is monophyletic. Figure adapted from Marc Ereshefsky's, Poverty of the Linnaean Hierarchy: A Philosophical Study of Biological Taxonomy (2007, p. 54) and The Reptile Database (Uetz, 2016).

One might look at the above figure and ask why this seemingly minute distinction matters. After all, the tree is essentially *the same*, regardless of whether or not you understand the class Reptilia to contain birds or not—these differing beliefs produce the exact same tree diagram. What *does* differ is the ways in which each school *interprets* this diagram to represent some argument about how animal groups form taxa—the basic levels of the biological kingdom that describe life on earth. And such taxa form the binding structural mechanism by which we communicate biodiversity information from person-to-person. Once taxa are listed in database environments, stating that these taxa occupy different class *positions* means they occupy vastly different spaces in the nested taxonomic hierarchy of the Catalogue (see Figure 17). In this figure an *evolutionary* point of view (according to Ereshefsky's model) situates birds (Aves) as a different class. Theoretical and interpretive distinctions amplify themselves in online taxonomic structures that represent relationships more rigidly than in traditional tree forms.

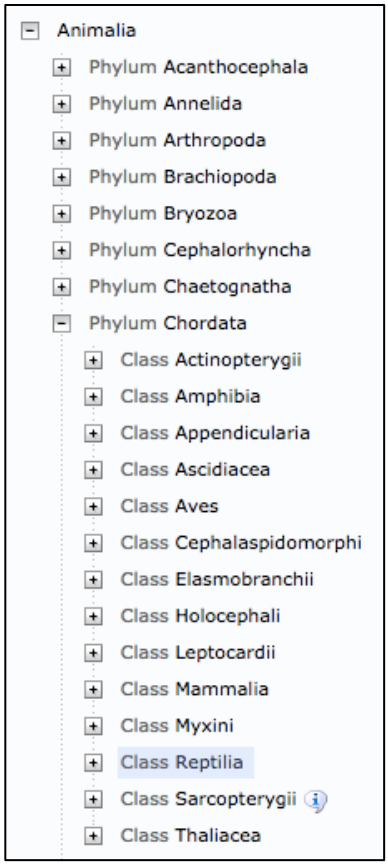


Figure 17. Catalogue of Life 2016 Annual Checklist taxonomic tree depicting the separate placement of the Class Aves from Class Reptilia in the tree structure (Species 2000, 2016d).

Cladistics, on the other hand—a school initiated by Willi Hennig in his publication, *Grundzüge einer Theorie der phylogenetischen Systematik* (1950)—in contrast to evolutionary taxonomy, does not accept paraphyly as part of their construction of taxa.[77] They would balk at separating *Aves* from the Reptilia group. Cladist's base their apportionment of taxa on the concept of genealogy and common ancestry:

> Cladists believe that classification should be strictly genealogical. However, membership in a paraphyletic taxon is not defined merely by common ancestry but also how much divergence has occurred among an ancestor's descendants. Because both of these factors are used for constructing paraphyletic taxa, cladists reject the existence of paraphyletic taxa. Membership in Reptilia, for example, requires being descended from a common ancestor. Yet according to evolutionary taxonomists, that common ancestry is not sufficient for membership in Reptilia: birds have significantly diverged from Reptiles, so they should be excluded from Reptilia. Cladists, on the other hand, deny the existence of the paraphyletic taxa Reptilia (Ereshefsky, 2007, p. 55).

[77] See also (W. Hennig, Davis, & Zangerl, 1999).

Due to the inference by evolutionary taxonomists that a genealogical relationship is not enough to define a taxa (namely, that *other* factors are taken into account, such as the previously mentioned "significant" changes that qualify as divergence into a new species), cladists reject the decision to separate birds from the Reptilia group. A major difference between evolutionary taxonomists and cladists, which lies at the heart of this essential conflict, is the extent to which evolutionary taxonomists and cladists understand, and choose to represent, the relationships between phylogeny and classifications. Every classification is but *one* way to represent a phylogeny that is altogether too complex to represent in any one graphical structure (D. L. Hull, 2001, p. 227)—there are far too many variables (known and unknown) involved in the process of evolution. As Hull states, "any one phylogeny can be classified legitimately in many different ways…only the most generic and impressionistic inferences about phylogeny can be drawn from an evolutionary classification" (2001, p. 227). For evolutionary taxonomists, this meant being willing to accept paraphyly as *one* way in which that complexity could be managed representationally via divergence.

Hennig chose "to represent only one—the sister group relation as it is exhibited in cladograms" (D. L. Hull, 1988, p. 131) in an effort to make the process of classification as "unambiguous" as possible:

> Two taxa are sister groups if they are more closely related to each other than either is to a third taxon. The evidence for this relationship is the presence of characters that the first two taxa possess but the third lacks—synapomorphies (D. L. Hull, 2001, p. 224).

Synapomorphies are traits derived from an ancestor and present in all taxa from that ancestor onward. This is in contrast to symplesiomorphies, which are characters that are found "in an ancestor and some but not all of its descendants" (Ereshefsky, 2007, p. 69). As Mishler reiterates,

> Hennig's central ontological advance was that homologous similarities are of two kinds, those due to recent, shared-derived homologies (synapomorphies) and those due to distant, shared-primitive homologies (symplesiomorphies). Only the former are useful for reconstructing the relative order of branching events in a system that is changing by descent with modification (Mishler, 2000, p. 662).

For example, if you look at Figure 16, above, *Aves* and *Crocodilia* actually share a more common *recent* ancestor, thus rejecting their relationship in the paraphyletic interpretation makes no sense according to Hennig's sister-group advancement. The result of this differing approach is that the tree diagrams each of these approaches produce say *qualitatively different* things and must be interpreted accordingly. Cladograms, such as those produced by Hennig and his adherents, "are not *phylogenetic trees*" (Ereshefsky, 2007, p. 71) in the formal sense. In a phylogenetic tree, the branching node represents a speciation event, while in the cladogram the branch represents the "joint possession of synapomorphy" (Eldredge & Cracraft, 1980, p. 212). The two trees between these different interpretive positions *mean completely different things entirely*. As noted by Hull, "The forks in cladograms do not represent species at all. Only terminal twigs in cladograms represent species" (D. L. Hull, 2001, p. 224). Further, true phylogenetic trees will also "depict genealogical history over time correctly" (Podani, 2013, p. 322), meaning that they have "additional information, in that edge lengths are drawn proportional to some attribute such as amount of change" (the evolutionary distance from one organism and another).

The final approach mentioned by Ereshefsky that that will be discussed here is phenetics, which produce a distinctly different representational hierarchical arrangement. In 1963, Peter Sneath and Robert Sokal popularized numerical taxonomy, a purportedly empirical method of constructing taxonomies based on statistical analysis of biological traits. Very much a classificatory method in reaction to "evolutionary taxonomy and other schools" (Ereshefsky, 2007, p. 60), numerical taxonomy, or phenetics, is a probabilistic method that clumps organisms together based on the general premise that those with the most phenotypic overlap are necessarily more closely related. Numerical analysis takes advantage of statistical and

computational methods to assess "taxonomic relationships … purely on the basis of the resemblances existing *now* in the material at hand" (Sneath & Sokal, 1973, p. 9). While evolutionary and cladistic methods were based on a mixture of common descent, homology, and the inference of the degree of change in a species over time, numerical taxonomy set out to produce an atheoretical mechanism for assessing relationships.[78] As articulated by Robert Sokal,

> Numerical taxonomists contend that evolutionary importance is undefinable and generally unknown and that no consistent scheme for weighting characters before undertaking a classification has yet been proposed. To weight characters on the basis of their ability to distinguish groups in a classification … is a logical fallacy. Since the purpose of employing the characters is to establish a classification, one cannot first assume what these classes are and then use them to measure the diagnostic weight of a character (1966).

Numerical taxonomy was touted as an empirical science, particularly because such analysis, in theory, would always produce the same result from laboratory to laboratory over a period of time (Sneath & Sokal, 1973, p. 11). Codes were created for each particular object trait: "'hairiness of a leaf' might be coded as follows: hairless: 0, regularly haired: 2, densely haired: 3" (Sokal, 1966, p. 114). Of course, as is the case with any method deemed empirical, one must critically assess the choice of variables or characteristics that undergo analysis for clustering. A notable weakness of numerical taxonomy is the fact that a relationship between organisms is defined solely by "similarity … operating on the assumption that the total phenotype accurately effects genotype. [Numerical systematists] believe that an unweighted measure of overall similarity provides an accurate determination of relationship. In so doing, pheneticists ignore the possibility of evolutionary convergence" (a circumstance where unrelated organisms evolve similar traits due to environmental influence) (Pietsch, 2015). Theory aside, it is important to the note the

[78] Acknowledging here that "atheoretical" means less about *not* holding a theoretical approach to classifying (which all scientific practitioners do), but more about trying to produce classifications whose interpretive, diagrammatic output, and object *positions* are not influenced by minute decisions influenced by a particular theoretical approach. Biological traits and parsimony-based analysis should dictate how organisms are depicted as related. Once variables and measurements were chosen, interpretive intervention was minimized. It goes without saying, however, that even phonetic positions are incredibly subjective in terms of how they identify variables, weigh them, which analysis they decide to employ, etc.

extent to which a numerical approach to classification would differ from an evolutionary or cladistic method of organization. As Sokal makes clear, "although close cladistics relationship implies close similarity this is not always the case" in numerical taxonomy (Sokal, 1966, p. 109). Phenograms (Figure 18) mapped out the operational taxonomic unit's (OTU) similarity, which could then be articulated as species concepts.

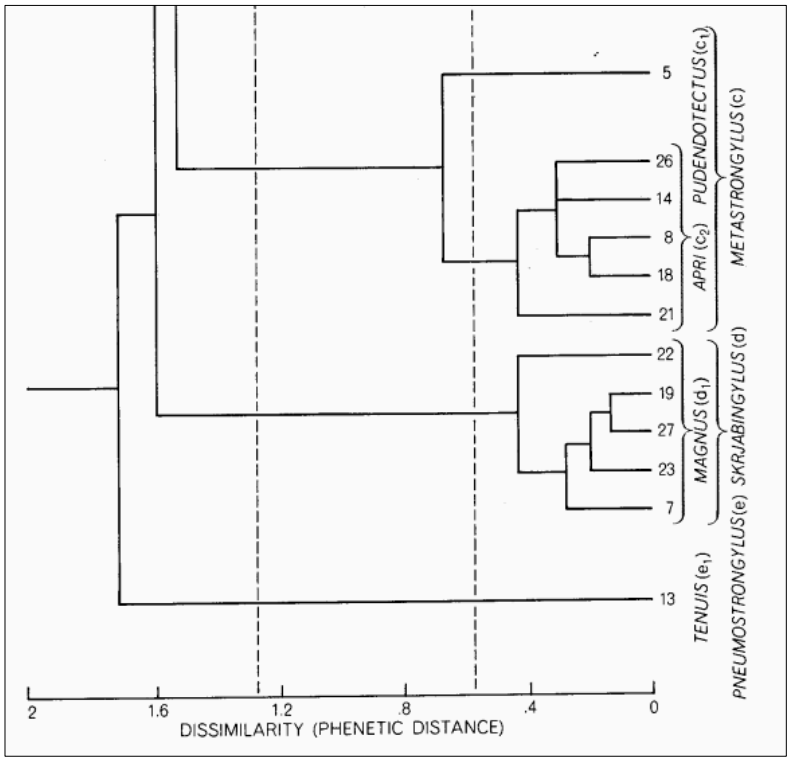


Figure 18. Cladogram. The output of numerical classification are phenograms such as this segment of a larger representation in Sokal's original *Scientific American* article (Sokal, 1966, p. 12). The numbers on the right represent species (OTUs) and their clustering is based on the similarity of characteristics between specimens.

While cladistics methods have become the more accepted standard in today's taxonomic environment (Dayrat, 2005, p. 407), there is still no universal consensus-based standard on which approach to take in practice. And there are still semblances of this typological thinking that continues into the present (as we saw in phonetics). As Doug Yanega indicated,

> Some people are more wired into [cladistics] than others; and mostly it hinges around parsimony. And not everybody buys into parsimony. And parsimony does not necessarily work for all types of data. That is why molecular people use different tools because molecular evolution doesn't necessarily have to work out to be

> parsimonious and then you also need to have your clock models built in for like the rates of gene substitution ... long branch attraction—those are problems that traditional cladistics doesn't deal with very well ... it comes down to how species are classified and how higher taxa are organized. A lot of traditional taxonomists, especially in Europe, they never really quite caught on to the whole bandwagon of doing phylogenies and understanding what monophyly and paraphyly were; they don't have a problem with paraphyletic groups. And you know, for a cladist, they would die in apoplexy if you try to make them accept a paraphyletic group. We still have people now that are happy to use paraphyletic groups and it drives other people crazy.
>
> Phylogenetics is evolutionary, it is not typological. There are the typologists, and that's the more traditional thing, and in a way we are almost coming full circle. Using barcodes for classification is getting very close to being typological about how your defining your [objects], and building trees by neighbor joining is very typological. And that's how this stuff worked when they first came up with numerical taxonomy and phenetic analysis. That was, you see which things are most similar to the others. And that's how a lot of the molecular barcoding stuff will work. And god, a lot of that is so much [like a] house of cards —you are trying to take the natural world and pigeon hole it in ways that are convenient and people see so many ways that they can do it and they don't always necessarily yield the same result (2016).

Regardless of which method of taxonomic arrangement prevails in our contemporary climate, elements of these schools demarcate sharp differences in the ways classifications are constructed and interpreted as modes of argumentation.

I want to emphasize that this admittedly reductive description of these taxonomic approaches is merely to make two basic points: that taxonomic approaches make (1) epistemological commitments to how they produce classifications—which is to say, that scientists make particular methodological decisions about what constitutes the creation of valid taxonomic knowledge and opinion; and (2) these decisions are based on various combinations of evidence that, in turn, force scientists to make particular classificatory commitments about how species *types* are fundamentally related in representational structures. These commitments permeate the entire structure of any given *produced* taxonomy. Taxonomies function on their own internal logic, and understanding the basic *positions* of objects within a taxonomy is dependent on the constellation of assumptions about what characteristics are used to define relationships, what theories inform the weight and importance of them, and how it is that the produced hierarchy should be interpreted as an argument. Classifications are not only theories

about how to create higher-level taxa; they are also diagrammatic representations that, in no way, can illustrate the true complexity of each approach.

The Catalogue, however, unlike particular descriptive-oriented based approaches, is not concerned with any one approach to producing taxonomic knowledge. In addressing this difference in the Catalogue of life, Michael A. Ruggiero, et. al. (2005), notes,

> Biological classifications can integrate diverse, character-based data in a phylogenetic framework, which allows a broad user community to utilize the disparate knowledge of shared biological properties of taxa. Phylogeny is, therefore, the basis for these biological classifications but there is still strong debate over their accounting for evolutionary divergence or information content other than the branching pattern [3]. Accordingly, classifications have often been labeled either phylogenetic or evolutionary, depending mainly upon whether or not they reject paraphyletic groups (2015a, p. 2).

The Catalogue must choose one management system to organize this taxonomic contingency. Let us now closely examine how a kind of consensus taxonomic structure is accomplished in this space.

Part III: Retrieval-Oriented Classifications*:* Toward a Consensus-Based Composite Instrument

I now want to shift focus and prioritize taxonomies, such as the Catalogue of Life, that serve primarily as *access* infrastructure. The underlying assumption of the Catalogue is that a data landscape that consists of multiple, fragmented taxonomies is a less than optimal *retrieval* circumstance. If, as Tony Rees indicates, a classification is intended "to provide a recognisable navigation structure so that persons entering the classification can (hopefully) find their way to their particular organisms of interest" (2009), then how is a user to *understand*, as Wilson makes quite clear, the instrument as a set of coherent relationships within the Catalogue? A user needs to *anticipate* the epistemological approaches to knowledge implemented and object positions within each of these systems in order to adequately harness their exploitative power. Knowing if a classification, taxonomy, or tree, was created using cladistics, evolutionary, or phenetic

approach certainly has an influence on how a user will navigate the system—and ultimately locate the particular position that taxa should occupy. The Catalogue, however, which consists of *many* taxonomies is a bit more complicated to understand in this sense.

For the Catalogue, coming up with a unified method of assessment for an innumerably diverse set of taxonomic productions seems like a nearly impossible task, despite Bliss's tendency to believe that equilibrium of knowledge is eventually reached. A model of the natural world is completely dependent upon the assumptions and metrics a taxonomist uses to make relationship judgments. As Kevin Thiele and David Yeates have stated, "Taxonomists routinely filter large arrays of observations through an extensive knowledge base to decide that this group of specimens or that set of prior taxa comprises a new taxon. Crucially, another taxonomist using the same knowledge base may validly arrive at quite different conclusions, and it may take time for thorough testing of the alternative concepts to arrive at *consensus* (itself subject to future challenge and refinement)" (my emphasis) (2002). *Consensus* is an appropriate word to use in this context, and one that merits more examination if we are to understand how *hybrid* systems such as the Catalogue of Life gain credibility (or not) as a taxonomic structure *independent of* the internally-consistent taxonomic opinion that we typically use to judge biological classifications (or classification in general). The key question becomes, what constitutes a consensus-based, valid representation of the external world? And, if universal classification (or unificationism, in Jonathan Furner's terminology (2013a)) is not the answer, then what might be a way forward to aggregate these disparate epistemological taxonomic approaches?

As we have seen, species concepts are contingent and fluid in their composition, and with such documentary contingency, it makes it difficult to define the relationships that connect them with a unified taxonomic instrument. David Hull, in *Science as Process* (1988), understands

scientific theories as kinds of types that themselves exist within particular historical lineages (1988, p. 515), prone to evolution and change over time. Kuhn (1996), too, acknowledges that competing paradigms (taxonomic or otherwise) were essential to structuring the punctuated normalcy of scientific practice. A. Broadfield understood consensus to be a core concept in the development of classifications, but one that was antithetical to the production of the scientific knowledge:

> But the scientific character of an age is a result of the way in which scientists classify their conceptions, not something to which they appeal in order to classify them. The active work of classifying scientific concepts cannot be an attempt at consistency with established knowledge, for to be established is to be almost dead, and long since classified…From the point of view of research, as soon as anything is agreed upon, it wanes, and 'agreed knowledge' is a signboard warning the seeker not to expect much in a territory whose possibilities have been largely exhausted. In [the hands of scientific workers] classification is a living study, and is part of their work of discovery…A scientist is often engrossed in and identified with a process on which he has been working on all his life…His interest is not so much in finding a ground of agreement with other workers on the comparative excellence of various processes, as in proving by demonstration of the superiority of his own reasonableness of descent. There would be no use in asking his opinion of the other processes, for he would say that of course their advantages were undeniable, but were outweighed by their disadvantageous" (1946, p. 72).

This tension between personal preference and the broader goal of facilitating data sharing between these competing taxonomic approaches is one that stands at the forefront of the Catalogue of Life's general efforts to consolidate taxonomic information. A lunchtime conversation with Thomas Orrell at the Smithsonian NMNH is a case in point, where we were discussing the importance of the Catalogue of Life as a taxonomic *management* tool. One of the biggest impediments to the ready communication of taxonomic knowledge, he articulated, is the adherence to taxonomic *traditions* as the *only* way to understand the organismic landscape. "Why does dogma overtake certain areas of the discipline?" Orrell asked, "Why is it that traditions can't be seen as separate from functional classification of information?" For example, he said, there should be one code [of nomenclature], but there are many. Yet people remain dedicated to keeping their own methods and ways; they are taught to understand the discipline in one particular way. "Why can't we agree upon a unitary classification? What is it that stops us

from doing that?" he further inquired. There are two issues of note to highlight in this brief conversation: first, one of the primary goals of the Catalogue is to provide a *unified* classification, not a universal one. The Catalogue wants to bring information together for access and rearticulation, *not* create a structure that has universal application in all circumstances. Second, the Catalogue is intended to provide a structure that can *communicate*, not (necessarily) one that argues a taxonomic opinion.

What Orrell is articulating, quite essentially, is that taxonomic classifications function on two very distinct professional levels: one that requires a taxonomy to facilitate data transfer *and* one that can (and must) be used to serve as the structure by which taxonomic arguments are made. How can we divorce the expectation that a taxonomy must be an opinion, from the idea that a taxonomy can serve primarily as a facilitator of information? Scientific communities (for any number of reasons) are divided as to whether or not such a goal can and should be met. The general methods by which classifications are conveyed in graphical forms (trees, cladograms, etc.), and the database environments that disseminate their information (browseable hierarchies, text searching mechanisms, etc.), give an *impression* of professional consensus that may not otherwise exist. What *kind* of consensus, then, does the Catalogue qualify as? And how can we understand such consensus to differ from the *kinds* of consensus necessary to create distinct internally-consistent taxonomic GSDs such as The Reptile Database (Uetz, 2016), or any other?

Indeed, Broadfield makes a sophisticated distinction that exemplifies Orrell's position: "to *arrive at* something by consensus is a rather different matter from organizing something in consistency with a consensus in which the logical order is supposed to be established, for the latter suggests a blind acceptance to dogma" (1946, p. 77). The Catalogue's function is not to organize a consensus of biodiversity theory to influence the approach to taxonomic construction

at a local level, only to articulate that consensus-based approaches to aggregating information are necessary, even if their functionality at a local level are not yet perfected. I certainly do not want to give the impression that taxonomic opinions such as those represented by Global Species Databases are "blind" in any sense of the term. My reasons for invoking this distinction made by Broadfield is merely to indicate that there are two ways in which we can conceive of *building a classification*: one that expects a certain degree of classificatory adherence or commitment, and one that expects *only* that a classification serve as a *functional* tool for communication, regardless of the classificatory consistency it embodies. I take the Catalogue to be the latter, based on a top-down approach, while taxonomic opinions to inhabit that former space are seen as a bottom-up method of constructing a classification. Rather than use the term *consensus* for infrastructures such as the Catalogue, which perhaps assumes *complete* coordination and general agreement within the domain on a *unified* structure and standard, I will use the word *composite*, which emphasizes more the production of a *complex*, compound structure ("composite, adj. and n.," 2016) with individual taxonomies forming the distinct parts of a larger taxonomic structure.

Articulating a composite taxonomic instrument.

> *["The Modernity of Classification"] interrogates the shift from classification-as-ontology, in which everything is defined as it is, to a more contemporary notion of classification-as-epistemology, in which everything is interpreted as it could be — or more precisely, the paper argues for a conceptual move from modern monistic ontology to late-modern pluralistic epistemological foundation for classification theory and practice.*
>
> *—Jens-Erik Mai*
> *'The Modernity of Classification" (2011, p. 711)*

I would now like to focus on the Catalogue's "access-oriented" classification specifically. Even as far back as Linnaeus, nomenclature and taxonomies were meant to improve information recall amid growing information stores (Müller-Wille & Charmantier, 2012) and to facilitate memorization of species (Ereshefsky, 2001, p. 366). This is to say that one way to look at taxonomies is that they are meant, primarily, to expedite information retrieval. Jonathan Furner

notes systems being judged as retrieval mechanisms should "enable access to documents in an *effective*, *efficient*, and *easy*" manner (Furner, 2009, p. 10). But even this is not so simple a task as it might initially seem in the Catalogue's instance. A classification system must be easily negotiated by the user, meaning that *documents* (which, in the case of biodiversity taxonomies, are species name-tokens that represent certain evidentiary documents that, together, constitute a concept) must generally be in a part of the taxonomy to be suitably assumed and locatable by a user.

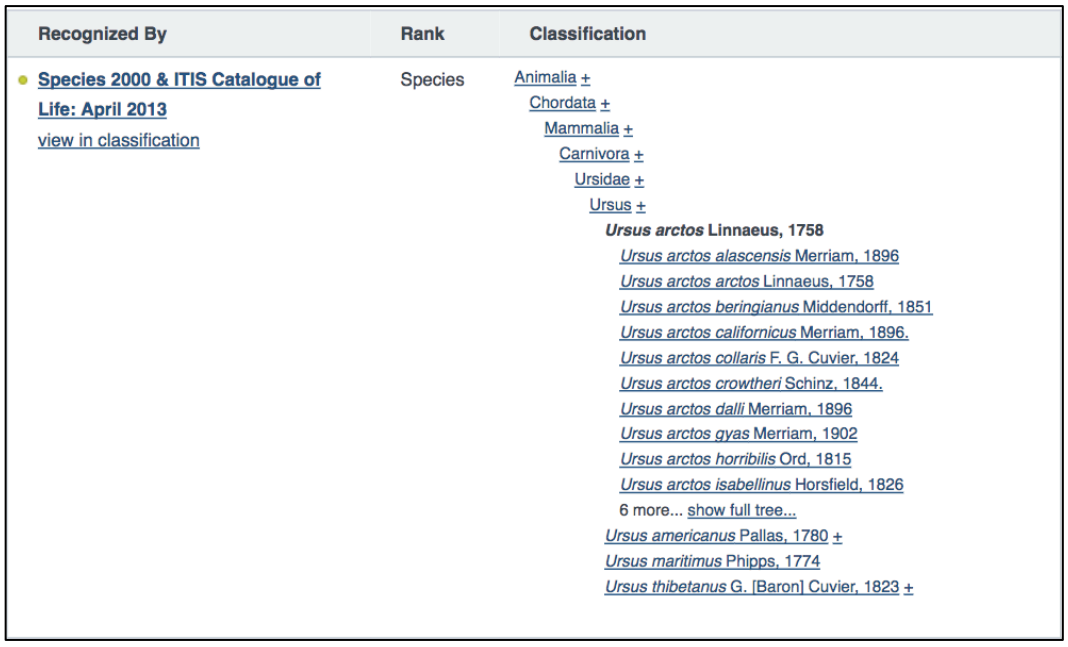
Referring back once again to our roadmap for this discussion, Bourgoin's "Global Names-Catalogue of Life Parameters" (Figure 12), we have finally made our way to the management classification of the Catalogue of Life, represented between Stage 4 and Stage 5. To recapitulate: we began at undifferentiated text strings, disambiguated into scientific names, found name-network relationships (historical, homotypic, etc.), and then assessed the constructions of "valid" taxa in classifications (necessarily produced by *expert* opinion and exemplified in the GSD repositories that make up the Catalogue). As GSD repositories come together, mechanisms need to be implemented to *integrate* these structures. If we think back to Wilson's differentiation between "The Catalogue" and "The Bibliographical Encyclopedic" instruments, the Catalogue of Life (despite its name) is less about a coherent taxonomic "subject" than it is an architecture that seeks to encyclopedically estimate "what the 'writing [or, this case, a taxonomic document] is good for'" (Wilson, 1968, p. 67) among a host of options within the documentary universe. Executive Editor Roskov said as much, describing his primary purpose as being to "[collate] the 'encyclopedia' chapter-by-chapter" as new information is added to the system (2016a). But there is not only one method to integrate and compile taxonomic information.

Integrative approaches.

The first task toward creating a composite structure for taxonomic coordination is to define what *kind* of integration a platform should optimally strive for. One way to represent and coordinate numerous taxonomies in a digital space is to present numerous taxonomic opinions on any one given page in an enumerative fashion. An example of this approach is the Encyclopedia of Life, which, as a reminder, strives to provide "information for every named species on the earth" through a one-page-per-species website ("What is EOL?," 2017). EoL identifies one taxonomic tree to display as part of its landing page, per species (Encyclopedia of Life, 2017c), which is chosen by a series of full curators that are either "credentialed professional scientists or EOL community members who have earned the respect of other curators through their work as assistant curators" ("EOL Curators," 2016). Acknowledging that there is no accepted hierarchy, EoL also presents a subsection of each species website that lists a number of other curated taxonomies that treats all these entities the same without a scale of value attached to them—an approach they call *taxonomic pluralism* (personal communication, June 15, 2016). There also exists a page that provides an even deeper level of "unvetted" *incidental taxonomies* that "help to bridge gaps in EOL coverage for groups that are poorly studied and not well represented in community maintained nomenclators" (Encyclopedia of Life, 2017b).

As an example, the species webpage for *Ursus arctos* (the brown bear) lists the "Species 2000 & ITIS Catalogue of Life: April 2013" taxonomy as its default taxonomy, selected by curator C. Michael Hogan (Encyclopedia of Life, 2017c). Six alternate, curated taxonomies are available that contain the species name *Ursus arctos*, including the Paleobiology Database, NCBI Taxonomy, the Integrated Taxonomic Information System (ITIS), and the IUCN Red List (Encyclopedia of Life, 2017b) (see Figure 19, below, for an illustration of the Catalogue and

ITIS hierarchies). Additionally, one can dig even deeper to see a total of fifty-two classifications that include representations from Wikipedia, as well as other (quite reputable) biological information sites (Encyclopedia of Life, 2017a). The sheer volume of taxonomies listed for *Ursus arctos* gives us a sense of how much complexity taxonomic databases like the Catalogue of Life are trying to control within its designated set of parameters. As of June 2016, plans were being discussed to implement *dynamic hierarchies* within the EoL interface, which would provide a mechanism for users to submit suggested changes given current taxonomic opinion—for both the variant taxonomies as well as the default taxonomy listed on the species home page of EoL. The goal here is to present the most *up-to-date* taxonomy for a given species at the cusp of cutting-edge scholarship. Given the constantly evolving nature of taxonomic opinion, however, it remains to be seen how this feedback mechanism will balance 'currency' with an interface that provides a relatively *consistent* browsing mechanism. Making all users happy with the chosen default taxonomic display, at any given moment, will be difficult to achieve, since current research is not typically universally agreed-upon.



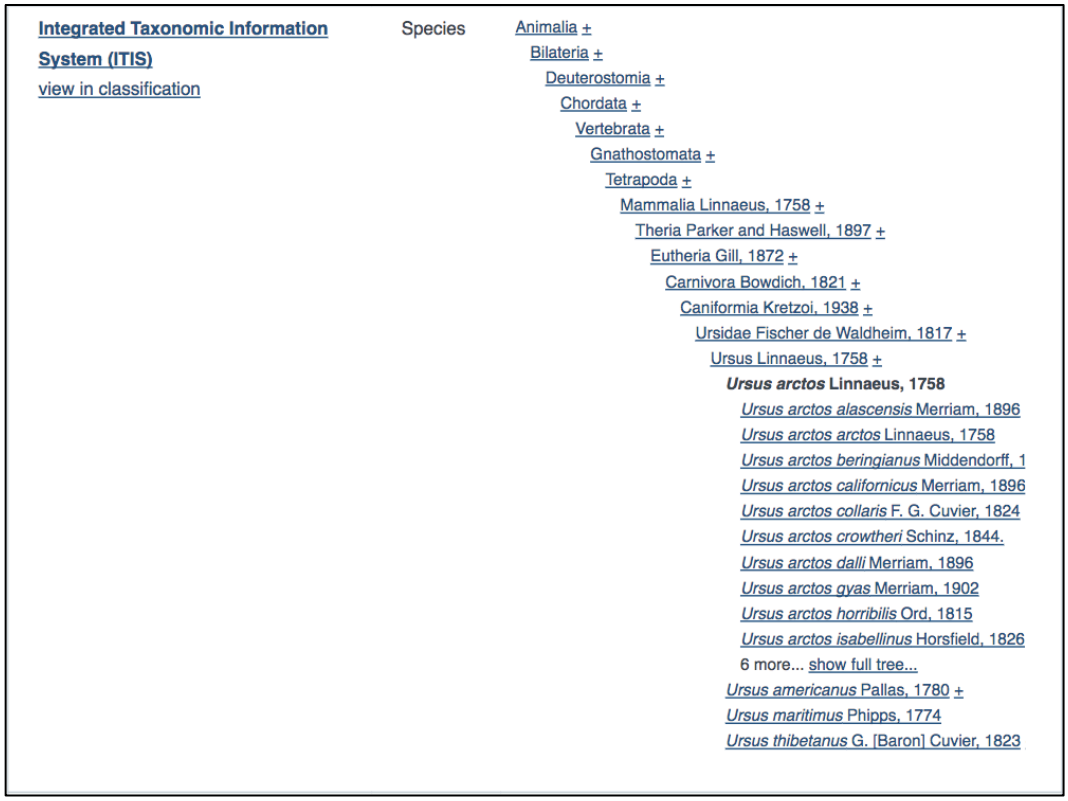


Figure 19. Two different curated taxonomies displayed by the Encyclopedia of Life for the species *Ursus arctos* (Encyclopedia of Life, 2017b). (Top) The (default) classification hierarchy for the species provided by the Catalogue of Life. (Bottom) The classification hierarchy provided by ITIS.

While the Encyclopedia of Life's taxonomic representations are usefully aggregated in this space, making it quite easy to compare them for differences, what this structure *does not* do is merge these various taxonomies to get a clear sense of the entire global diversity of life in one coherent system. One reason for this is quite simple: EoL is not a taxonomic database, *per se*, in that taxonomy is not its primary focus; rather, *species*-level knowledge is what is emphasized here, and the collocation of data generated around the world at this level. While the organization of the EoL pages is certainly organized by some back-end taxonomic system, that system in-and-of itself is not front-and-center as part of the browsing experience. And when users *do* need to browse, they have multiple taxonomic options by which to seek out information. Thus, EoL expresses what I call a *divided* plurality: the comingling of multiple taxonomies side-by-side to allow for a user to choose that taxonomy most suited to their purposes. The taxonomies

themselves remain intact as they are created at the source database level. And while the taxonomies technically constitute the "backbone" of the species in the EoL, because those taxonomies are always shifting (due to curatorial intervention), any given species may change position over a period of time. This makes it incredibly difficult for non-expert users of the site (Species 2000, 2016c).

Hierarchies in hierarchies: "Ornaments on a tree."

In order to circumvent the unavoidable and irreconcilable issues that arise from these different taxonomic factions, the Catalogue has a team of experts that have decided upon some basic formats that structure all incoming taxonomic data purely on the assumption that taxonomic compromise is necessary for data collection and reuse. Choosing between multiple taxonomies, as you do with EoL, is a virtue in one sense: you get an impression of the overwhelming variance in expert opinion, and thus you have the potential to make a more informed choice as to how you wish to experience the interspecies information flow. Discriminating between a *more*- and *less*-*optimal*, situationally-relevant (Wilson, 1968, 1973) arrangement for most individuals, however, requires scientific knowledge that most individuals do not have. In order to facilitate the long-term data management of species-level data, reduce taxonomic redundancy across the biodiversity landscape, and broaden the user base of system that requires expert differentiation of multiple taxonomic trees, a new mode of taxonomic arrangement was needed to enable communication among these many fragmented systems. Adherence to taxonomic opinions based on internally-coherent hierarchies will never be able to serve this purpose, plastic and idiosyncratic as they are as scientist's opinions changes over time. "The phylogenetic and classification work is always shifting and it is different *kind* of work than that of trying to *communicate*" (T. Orrell, personal communication, June 15, 2016).

The Catalogue of Life staff, thus, took it upon themselves to try to bring a sense of taxonomic stability to the biodiversity infrastructure/iLife consortium by implementing a *management hierarchy*, specifically built as a data communication mechanism to make organizational control of intellectual (documentary) assets more cohesive. The management hierarchy merges multiple taxonomies into one functional system. Unlike EoL, where *divided plurality* was defined by the *choice* between many options, the Catalogue's *integrated plurality* comingles many hierarchies in one system. The Catalogue's founder, Frank Bisby, believed that, in order to truly understand the extent of biodiversity knowledge, there needed to be a mechanism to aggregate data in a unified and *coherent* manner. Then, and only then, would scientists be able to address gaps in species knowledge on a global level. Thus, the Catalogue has had a broader mandate than most taxonomies from its inception: rather than argue a specific methodological structure, this taxonomy would chronicle the breadth of extant taxonomic knowledge. The Catalogue's Executive Editor conveyed,

> We were firmly dedicated to the position that we should provide users [with] a single view of taxonomy. A single, simplified, unified view. Unified through different codes—zoological, botanical, bacterial codes of nomenclature. Frank [Bisby's] idea was to make a single index. A kind of *Yellow Pages*, if you like. A user could locate a name and … look where that name is in the classification, [find] synonyms, accepted valid names, etc., and where species [are] placed within a genus, family, and so on (Roskov, 2016a).

While the Catalogue began as only a list, it soon became apparent that the list format overlooked the more holistic view necessary to gauge what biodiversity segments were least documented and studied. But how was this aggregation to be achieved in a practical sense? Merely bringing together datasets and placing them side-by-side would leave a great deal of overlaps and redundancies between taxonomies, along with juxtaposing organizational schema that conflicted on some basic structural levels. Adding even more complexity to this task, not all databases included the appropriate higher-level taxa categories (for example, above *order*) necessary to easily assemble the taxonomic 'blocks.' Scientists who build their own databases often have no

need to include higher ranks since *they* built it and thus had no need to indicate what, to them, was implicit knowledge about their particular species data set. The Catalogue's *management classification*, then, was created to provide a basic organizing structure for these taxonomic misalignments. The management hierarchy "disjoints" taxonomic practice "from the desk" of individual scientists (Schalk, 2016a).

The management hierarchy uses the "standard formal [higher] categories" of the Linnaean-based ranking system (phylum, class, order, family, etc.) most suited to accommodate the largest amount of taxonomic information derived from contributing databases. "The classification recognizes two superkingdoms, Eukaryota and Prokaryota, and seven kingdoms: Animalia, Archaea, Bacteria, Chromista, Fungi, Plants, and Protozoa. The classification also includes 1,467 orders of living organisms in 351 classes. Where certain taxonomic associations are still unresolved, the panel provides an interim recommendation" (Species 2000, 2016c). The Catalogue's ultimate goal,

> Is to provide a hierarchical classification … that (a) is ranked to encompass ordinal-level taxa to facilitate a seamless import of contributing databases; (b) serves the needs of the diverse public-domain user community, most of whom are familiar with the Linnaean conceptual system of ordering taxon relationships; and (c) is likely to be more or less stable for the next five years…Beyond the immediate use for the CoL, the hierarchy is valuable as a reference for taxonomic and biodiversity research, as a tool for societal communication, and as a stable "classificatory" backbone for biodiversity databases, museum collections, libraries, and textbooks, to name a few applications (Ruggiero et al., 2015a, p. 2).

Like ornaments on a holiday tree, the Catalogue of Life attaches GSDs to the management classification above the node of attachment of each database (See Figure 20, below) (2015a; Gordon, 2009). From this node downward, the general composition of the GSD is maintained in its entirety—the expert (that provides the GSD) serves as the final authority for its composition. Each species database sector is linked only at one node in the classification. "The taxonomic rank of the highest taxon at this attachment node varies from one GSD to another," depending on how the Catalogue's editor sees this source database interacting with the other adjacent

taxonomic entities "(e.g. [the] Conifer Database is attached [at the] phylum [level], [the] sector of Cercopoidea Organised Online is attached as superfamily [level], [and the] sector of [the] ILDIS World Database of Legumes is attached as one family)" (Renear, Sacchi, & Wickett, 2010, p. 5). As part of the Catalogue's standard dataset (Species 2000, 2014), the GSDs are required to provide data in fields that represent their particular chosen hierarchy and taxa, including the highest taxon covered, as well as all of the taxon beneath that highest taxon level.[79]

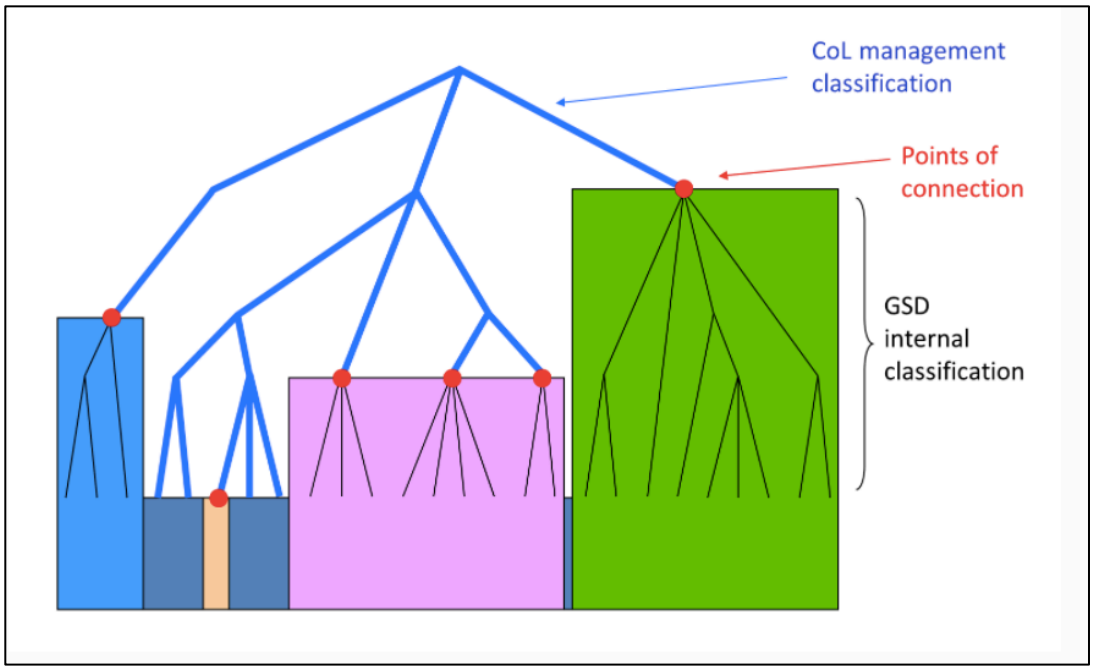


Figure 20. Schematic of the Catalogue of Life Management Hierarchy interacting with the GSD Internal Classifications. Original image label: "The Catalogue of Life retains the GSD's own classification below points of connection and uses the management classification above" (Species 2000, 2016c).

The process of creating the management hierarchy was a difficult and complicated one, full of a great deal of heated debate and deliberation—most of which was accomplished through email and other remote mechanisms (T. Orrell, personal communication, June 15-16, 2016). As part of the process of finalizing the ultimate taxa categories for the management classification, "authors consulted more than 200 sources, most of which were from recent taxonomic

[79] See Figure 8 in chapter two for the standard dataset fields.

publications and websites" (Ruggiero et al., 2015b). As was conveyed by Peter Schalk, "The first hierarchy was replaced about 3 years ago … after 2 years of work [between] quite a few players in the Species 2000 field … It's still not perfect, but good enough...and even then you still have conflicts because their vision is not always the same" (2016b). Issues central to these deliberations were how taxonomic authorities were to be set, and how one taxonomic view was to be adopted over any other. Striking a balance between three basic user groups—experts, institutions, and non-scientists—is about negotiating which classification virtues are worth maintaining and which are suited for only professional conditions. In the end, the Catalogue is about managing data, not about managing internally-consistent taxonomic opinion, and thus, this hybrid structure was implemented to facilitate ready access to as much information as technically possible, while always deferring to taxonomic opinion as the optimal result in subsections of the Catalogue where appropriate.

But, as Paul Kirk, of the Royal Botanic Gardens, Kew, made me aware of, it is important to distinguish between two distinct entities and initiatives happening in parallel within the Catalogue infrastructure:

> The Catalogue of Life has two elements: those people connected with the Catalogue of Life that are producing a higher-level management hierarchy for life on earth ... [Thomas] Cavalier-Smith, Dennis Gordon, the ITIS people, myself, AlgaeBase ... they are looking at what's out there and trying to come up with a *consensus* for life on earth.[80] That is happening in parallel with the management hierarchy that used to manage the content in the Catalogue of Life, which is slightly different. What the Catalogue of Life decided to do early on was decide that that we are not trying to resolve all issues connected with classification. If somebody comes to us with a GSD for a family or a higher rank, we're going to adopt their classification and plug it in, as far as possible, with existing classifications and try to avoid any overlaps So we've got these two hierarchies: [one with] everything as we see it down to ordinal level (not to family or genus)—a group that is producing higher-level stuff—but that might be out of sync with what the people are plugging things into. And everything is moving forward, but within the Catalogue of Life it is moving forward on two fronts. (P. Kirk, personal communication, August 30, 2016).

Kirk articulates this distinction to emphasize the difference between the theoretical work going on to produce a consensus system to organize all life that must negotiate many taxonomic

[80] See (Ruggiero et al., 2015a).

traditions (the higher-level management hierarchy as described above), and the more *practical* work of creating the *actual* taxonomic system that is a composite structure fundamentally altered by the ingest of GSDs from all over the world (the Catalogue of Life as a dynamic database). The former has a stability factor of about 5 years (the general timeframe the Catalogue global team task force is set to review the management classification structure), while the latter, practical management taxonomy changes along with the ingest of every database that enters the Catalogue of Life taxonomic backbone.[81]

In some ways, Kirk is pointing to the difference between the *work* entity of the Catalogue—the idealized notion of what a *full* management hierarchy might look like—as well as the *text* of the Catalogue—the species token and taxonomic relationships that constitute the Catalogue as part of monthly and annual editions. The *practical* consensus system is managed by a team of editors, led by the Executive Editor, Yuri Roskov, at the Prairie Research Institute at the University of Illinois, Urbana-Champaign. Roskov handles all of the decisions regarding data transformation, and maintains contact with the various data providers that contribute GSD systems to the taxonomy. Part of this communication involves negotiating with database managers who contribute to the Catalogue free of charge as an open source piece of software.[82] Additionally, Roskov has between seven and ten individuals serving on a board of editors that help him negotiate with regional hubs and umbrella providers such as those located in Europe,

[81] See chapter two for information on publication timeframes for the Catalogue, which will correlate with the frequency at which the consensus management classification also changes.

[82] As we have seen thus far, taxonomies represent the output of real scientific work and hypothesization, so asking a scientist to contribute data is asking them to contribute many years of effort—potentially the result of an entire career's work. Further, biodiversity database maintenance is a costly endeavor, rarely subsidized by the institutions where the scientists are employed. Concerns raised by scientists contributing data include, not getting properly credited by those that download and use data; not being able to control manipulation of their data once it is downloaded by other parties; and not having full control over data transformation as it enters the Catalogue of Life. Access agreements that attempt to clarify and reconcile these issues are, of course, negotiated, but within a distributed database environment such as the iLife consortium, full control of data throughout its full lifecycle is nearly impossible.

China, or South America (2016b). The intellectual and theoretical work of building the management system is designed to provide flexibility to the practical work of building the Catalogue's database taxonomy.

A key issue to note is that, as discussed above, on a very practical level, each of these different database 'ornaments' has a classificatory logic that may differ from one another. As we have indicated, not everyone organizes animals or plants or bacteria in the same way. It makes sense to now think about how editors choose databases for ingest, and how their interference might be reconciled in the database space.

Discriminating taxonomic contributions and filling gaps.

Consensus is a contentious word within the context of the Catalogue, for the various camps within the practice of taxonomy cannot (and by definition, should not, given that taxonomy is a science) agree on o*ne* global standard for constructing hierarchical relationships. In the biodiversity database world, at least for the purposes of the Catalogue of Life, assessment of taxonomies has nothing to do with a correspondence to some *a priori* notion of *reality* or with what specific method of taxonomic opinion is invoked, but rather is based on the general notion that it is *good enough* given the state of biodiversity science and the reputation of the contributing individual at hand. In general, while the Catalogue is "reflective of [phylogenetic]" (Ruggiero et al., 2015a, p. 2) approaches within its hierarchical structure, it does not purport to be a cohesive phylogenic tree. Rather than method, *reputation* is considered the hallmark of good contributing GSD classifications:

> No, it doesn't matter to me at all [how they build their classification]. First of all, I am asking whether this taxon is covered globally or not in their resource, so whether it's a monograph or whether it's partial just for a region, or for an ecological niche (say, marine), and if they say, yes it is covered globally and it is produced to the best of our knowledge of [the] taxa, and has all modern publications, that is good enough for me. And, of course, if they say it follows the most up to date classification. There is nothing like a conversion method for different classification specialists. So they might be different. And for me, it means nothing, because taxonomists may never come to agreement. For example, in the scorpion group, there are

> two big groups whose opinions conflict with each other and this can't be stopped. And actually, my honest belief is that we should not restrict diversity of classification hypotheses. So users may, or should, have a choice of the arrangement that is best for them. In the Catalogue of Life, following the approach of a single classification, we are not saying our classification is right and prefect, but we are saying we make our choice (Roskov, 2017).

In an environment where taxonomy *cannot* be pinned down and labeled correctly, all the Catalogue has to go on is whether or not a scientist has done due theoretical and methodological diligence in producing their classifications.

And while most classifications *tend* to use similar methods, minute differences cause misalignments that present themselves as gaps and overlaps. Moving taxa from one area of the taxonomic tree to another is not a simple editorial task. Recall the process of nomenclature and how, to a certain extent, nomenclature reflects taxonomic *positions* (the first part of a name indicates genus, the second part species). If a species is moved into another genus, the first part of the binomial name needs to be changed to reflect this modification. If the Catalogue were to attempt to do so, the new name would qualify as a 'new combination' (Global Names Architecture, 2017d) and, in order to be accepted by the taxonomic community as a valid nomenclatural act, it would need to be published in accordance with the particular code governing that group. Aside from causing undue confusion in the taxonomic community (changing a name for *merely* technical purposes rather than for descriptive purposes would be antithetical to normal nomenclatural practice), the process would be incredibly time consuming and expensive. Taxonomic practice occasionally interferes with the overall technical mission of the Catalogue.

There is also a fundamental tension between GSD authority and editorial review. While the Executive Editor is a taxonomist by training, no one person can be a specialist in all areas. As Roskov remarked, "We are quite limited in our ability to control this data because the taxonomic

databases that contribute their data are the "kings in their kingdom"—they are responsible for taxonomic visions and concepts in their sector" (2017).

Even as the Catalogue has taken shape over time, gaps inevitably present themselves in the management taxonomy. More charismatic species—those species that get more attention by scientists and policy makers (Bowker, 2008, p. 146)—have a tendency to get described and classified more exhaustively. Certain species groups, on the other hand, such as worms (*Annelida*) and mollusks (Kunze, Didžiulis, & Roskov, 2013), have a great deal of descriptive gaps in the literature (2016c). Information on these species just is not yet available in GSDs. There are also circumstances when GSDs are precluded from being included in the Catalogue due to user agreement disagreements on intellectual property issues. Databases are the result of difficult, individual labor, and, at the local level, one person (or group) controls the description and data related to one entire taxa; and so if that *one* individual refuses to contribute, an entire segment of the tree of life is excluded from the Catalogue's global taxonomy.

In order to fill these gaps, the Catalogue of Life attempted to implement an *intermediate* taxonomic space that they called proto-GSDs. Taking regional species databases (RSD) as their core dataset, proto-GSDs attempt to fill gaps that global species databases could not fill. Recall that RSDs are regionally specific, covering broad geographic spaces that can sometimes be as large as the country of China. Given this broad coverage, they often overlap with GSD content—and, in fact, some of the collected GSDs are actually *also* part of RSDs. For example, an RSD like the World Register of Marine Species (WoRMS) (2017b) aggregates GSDs in a similar manner as the Catalogue (and in some cases, also produce their own taxonomic backbone). Unlike GSDs that are often organized around specific *taxa* and have *global* coverage, RSDs are

more difficult to integrate given their multi-taxa and geographically specific composition.[83] Editors will often take chunks of RSDs and fill in taxa that are not adequately covered in the GSD dominated Catalogue taxonomy. But RSDs have one significant downside: they do not necessarily cover a given taxa globally.[84] Proto-GSDs, then, were meant to compensate for this issue, by artificially creating a GSD taxa space using multiple RSD segments compiled into one taxonomy. As explained on the Catalogue of Life Blog,

> A proto-GSD combines multiple regional checklists to try to achieve greater coverage for a particular taxon. One would think this would be quite straightforward, after all isn't it just a case of combining datasets and removing the duplicates? Well not exactly, as combining regional datasets can lead to all sorts of taxonomic issues because of possible duplication in species names and also the conflicting species concepts and classification systems. Take for example the Family Gentianaceae in the Plant Kingdom. This family of plants has an estimated 1650+ species worldwide. We have two current suppliers of Gentianaceae to the Catalogue of Life - ITIS Regional database and Catalogue of Life China. Together they supply the Catalogue of Life with 552 species in addition to 82 infraspecific taxa. The species Gentianella acuta (Michx.) Hultén appears in both checklists where it is a synonym in ITIS Regional and an accepted name in the Catalogue of Life China. This is because some of the species of Gentianaceae are cosmopolitan (ie present in North America and China) and the taxonomic concept (ie accepted name or synonym) is different. To combine the datasets the Catalogue of Life editors had to resolve these issues before publishing it in the Catalogue of Life. Gentianaceae is now part of a 'proto-GSD' (Matthias, 2013).

No specialized software or interface was created for the Catalogue to perform these aggregating tasks, thus all of the editing work was being done directly within the database using .csv and Microsoft Access files (Roskov, 2016c). Designed software could potentially flag taxonomic conflicts and recommend master classifications between overlapping structures, thus releasing this burden from the taxonomy editors. Despite early attempts at building this infrastructure, such software mechanisms never materialized. One of Roskov's major tasks as Executive Editor is to manually build proto-GSDs to the best of his taxonomic ability. Once compiled, these databases are tentatively placed in the Catalogue to bridge existing gaps (knowing that the "aggregation" is "primitive" in nature and that it is not—yet—scrutinized by a recognized expert

[83] In the long term, the boundaries of nation and regions change for various political and economic reasons, which could potentially pose the problem of how to assess the *coverage* of RSD databases.

[84] Some RSDs like WoRMS and ITIS *do* cover species globally, but others, such as the Catalogue of Life China, do not.

in that taxon) (Roskov, 2016c). This kind of manual intervention, however, is costly, and the Catalogue is unable to revisit these proto-GSDs for each Annual version produced by the Catalogue. Until software is available that can mechanize this process, the temporary gap measures will have to suffice.[85]

Despite these proto-GSD efforts, there is no aggregative practice that is generally accepted by taxonomic specialists. "Professional taxonomists do not appreciate any kind of technical exercises like I described," Roskov indicates, "For example, a plant list that is being built using software by [the] Missouri Botanical Gardens ... where they are trying to merge regional floras. I spoke to professional taxonomists ...and they are very much skeptical about this work. It is a very political process."(2017). For the Catalogue, such interventions are better than nothing. "A draft checklist is better than [one with] gaps," Roskov indicated. If the goal is to get an up-to-date snapshot of the world's biodiversity in one coherent structure, the Catalogue will do all that it can to achieve this coverage, even if the mechanisms are temporary and imperfect. A major difference between *taxonomic opinion* and *management hierarchies* is this ability to accept flexibility and contingency. The Catalogue's taxonomy is a practical tool that understands and embraces its limitations. As I was told on numerous occasions: it is better to embrace taxonomic disagreement than to wait for a professional taxonomic consensus that will never materialize.

Part IV: Extensive Flexibility: Broadening Wilson's Schematic

What, then, might we be able to say about the *instrumentation* of the Catalogue versus the instrumentation of any other taxonomy, consistent as they are in their internal structure, classificatory relationships, and ontological stability? One unique aspect of the Catalogue's

[85] See chapter five for discussion about GBIF's taxonomy building algorithm.

taxonomy is the extent to which it, as a system of organization, is not *only* an information structure by which *represented documents* are placed in particular categories and relationships, but it is also a complete *taxonomic document* that is subsequently *extended* in different systems toward different purposes. In looking back to Patrick Wilson's, *Two Kinds of Power* (1968), he presented two powers that exist within the bibliographical universe: that of *descriptive* power, defined as the ability to describe documents in such a way that we can call-up a set of undifferentiated documents; and *exploitative* power, which he describes as the ability for the *use* of texts in a manner most relevant to the circumstance at hand. A trait of both powers is that they are focused on the documents *within the system*. In other words, Wilson's approach tells us how to describe objects in such a way as to provide the best use of *them*—and only them. And by extension, this would create a more powerful and functional bibliographic instrument.

The instrument itself in Wilson's narrative—the catalogue, or whatever organizational structure provided to seek out documents—is understood to remain intact as a cohesive entity. Wilson did not include (and perhaps couldn't anticipate given the primitive computational abilities in 1968) in his analysis the possibility that a given system could be made with the *intention* of being repurposed, recombined, and redefined in a plethora of other infrastructures (within the iLife consortium and beyond), through a variety of means. This seems to me appreciably different than the expectations placed on other classification systems that we typically see in the documentary and bibliographic space. This is a kind of control over the *system* represented by the Catalogue of Life that is not covered in Wilson's schematic. But this *control* is not a control at all in the traditional sense; it is the ability to *forfeit* control of what, in traditional biodiversity taxonomic circles, is highly valued: the internal coherence of taxonomic opinion in deference to a gain in *exploitative power*. In other words, in order to increase the

exploitative power of both the represented and recorded documents (species representations and their linked referents), as well as the *emergent* taxonomic document itself, a modicum of structural flexibility is necessary in the domain of *descriptive control*. Objects within biodiversity taxonomies need to be organized, related, and structured in such a way as to *make* them flexible. This is an *active* process, not a process that just happens to the Catalogue. Interestingly, the Catalogue's approach to biodiversity taxonomic organization, makes the case that this extensive flexibility actually increases the *exploitative* power of the Catalogue's document, given that it provides a mechanism for global documentary access in a multitude of different circumstances.

Such systemic flexibility is what defines the Catalogue: Catalogue staff create a consensus-based composite taxonomic structure that is meant not only to provide a means to both classify and organize documents, but also to provide a structural backbone in other systems that, in practice, rearticulates and re-presents the Catalogue's classification according to their various classificatory and epistemological interpretations. The third function of the Catalogue (aside from a species list and a taxonomy) is to be an instrument *of* organized knowledge (as opposed to an instrument *to* organize knowledge). One might argue the position that such re-use of a classification can occur with *any* classification. Somebody, can, for example, take the Library of Congress schedule and implement it for their own local purposes (whether or not permission was granted or needed), manipulating it as the see fit.[86] And this can certainly be seen as a kind of *imposed* power onto an instrument or system. But differentiating this kind of *imposed* use as categorically different from the use of *any* document for any particular purpose

[86] Of course, the point of the Library of Congress classification—or any bibliographic classification—is its implementation in many local repositories. However, the LoC system is *not* downloaded and then amended to local circumstances—this would defeat the purpose of using the system since you would never be able to use the main classification schedule again without causing a great deal of headaches.

would be a fruitless enterprise. An *imposed* repurposing is not unique to the Catalogue or any other *instrument*, so it is not a power in Wilson's sense. Wilson's entire discourse in *Two Kinds of Power* aims to discuss how it is that we can *control* a certain set of documents or texts systemically so that they can be best *described* and *exploited* within a documentary (or in Wilson's term, bibliographic) universe. One of the Catalogue's central aims is that it is *designed* to be extended into other digital and organizational domains. The structures, policies, and standards that the Catalogue has imposed make such reuse possible in multiple ways. As do the mechanisms to obtain feedback and input on the recursive updating of that document as new Annual editions are released.

The concept of *extensive power* helps us understand the potential functionalities taxonomies have within digital spaces. It also provides us with a mechanism by which we can conceptualize classifications without fully internal and classificatory coherence, which nonetheless serve an integral and powerful mediative purpose in the biodiversity domain. Among biodiversity infrastructures, such as those in the iLife ecology, such an extension is an essential part of what makes this and other infrastructures work—and how it is that scientists can produce functional mechanisms for coordinating local knowledge in global spaces.

Conclusion

The biodiversity taxonomic instrument is a complex machine. A fundamental question that arises within the space of eScience ecologies such as the Catalogue of Life is: Is it more important that these classifications be scientifically correct (as in, for example, representing current phylogenetic assemblages) or to more easily facilitate data sharing. And this tension is strong. I want to reiterate here that there is a sense in which the catalogue, and taxonomies in general, are both heuristic spaces (idiosyncratic, hermeneutic) *and* information communication

systems (just getting data to go where it needs to go). The vibrant conversations that went on in some of the biodiversity meetings I attended illustrate how strongly people's professional identities are tied to what we see as a very simple tree structure. But these two applications (heuristic usage and data facilitation) have very different purposes within scientific practice; and, really, they depend upon each other to create a robust and functional scientific production system. These are entirely different user groups whose needs and concerns the Catalogue must strive to balance going forward. And I mean technically, but also socially: how do you get buy-in from scientists who do not want their hard work transformed just because a data management structure demands it to function globally?

Whenever I think about the potential *powers* of the classificatory space I recall my aforementioned meeting with Tom Orrell where he asked, "Why does dogma overtake certain areas of the discipline?" It seems to me that dogma extends far beyond epistemological and methodological positions to build a taxonomy, but it also broadens to the ways in which we conceptualize the consistency and *authority* of classificatory spaces. Hope Olson (2002) helped us see through these authoritative structures and challenged us to push against them by imagining alternative techniques. Thinking of the Catalogue, what can be more alternative than setting aside one of the fundamental tenets of classification: control over internal consistency? Not only is the Catalogue composite in nature—that is to say, it is a database meant to aggregate many contributed database taxonomies—but it is also *consensus*-based, meaning that the taxonomy itself is *a self-recognized and designed negotiated object*. The Catalogue recognizes the contingent nature of *all* documents (species documents, nomenclature, and taxonomies) in the biodiversity space and is striving to make a space where that contingency can find some usability and functionality on a larger scale.

This prompts a larger question to the Information Studies community at large: what is a *good enough* classification system for the documents we describe and organize? And how can such consensus-based negotiations help diversify our systems, and share and collocate information more effectively and flexibly? To a certain extent I see composite structures like the Catalogue, exposing the fact that documentary systems, organized as they are with a multitude of classificatory schema, have *never* been completely consistent in practice or free of their own individuated forms of dogma—we see this in the plethora of classificatory approaches, theories and methodologies to constructing and hypothesizing organismic relationships. A. Broadfield once wrote,

> A common dictum is that classification should not be critical. Whatever precautions a classification may take, it will be critical. For it is a system of expressed judgments … The endeavor to avoid criticism can have only one result—the monumentalizing of beliefs which are looked upon, wittingly or otherwise, as being beyond criticism. Thus an attempt to achieve impartiality can become an insidious form of dogmatism (1946, p. 78).

Critical views of the Catalogue stem from its designed flexibility and from the editor's ability to acknowledge classifications as what they are: constructed points of view. Dogma takes on a different guise in this space: the only thing that is true is that the taxonomic construct *will* change. Method and theory are understood quite differently here. Embracing a pluralistic approach has been a key component for the scientists involved with the Catalogue: "[the] actual complexities of phylogenetic history emphasize that classification is a pragmatic human enterprise where compromises must be made" (Adl et al., 2005, p. 4). To embrace a plurality and to push against the ongoing debates about what constitutes a *right* and *wrong* mechanism for the organization of documents, seems to me a much more practical way of understanding the potentiality of classificatory spaces. It also shows the potential of embracing the critical nature of taxonomic spaces in general. If documents are contingent, then the classificatory document is no less stable; to think otherwise is to live within the simulacrum of control. Far worse in practice,

perhaps, is that if we assume systems to be at their core, coherent, then we begin the exploitation of a system at a *power* far less than we could otherwise have if we acknowledged (and could assess) their inherent, built-in limitations. The extent of our *potential power*—exploitative, descriptive, or extensive—over or through a system is only limited by the knowledge we have of the *instrument* in question. The fundamental relationships that create the paradigms we function within—implicit or explicit—need to be openly criticized and explored. We need to reach for Hope's eccentric techniques. The Catalogue, I think, helps us see both these tensions and this capacity.

One of the potentialities created by a 'knowledge base' like the Catalogue is that it can be recombined to create new forms of knowledge, with the ultimate aim of understanding the historical development of taxonomic knowledge and evolution. This aim has been met with mixed reactions. Moving onward, the next chapter will more closely examine the possibilities and critiques of the Catalogue. Can the Catalogue be used to answer broader evolutionary-based questions? In what ways is the Catalogue changing taxonomies that it brings into its structure, and how is the Catalogue changing *itself* as part of this process? And finally, what specific critiques of the Catalogue have been expressed by professionals in biodiversity practice—most of which are equally invested in the overall aim of information communication and centralization of services?

5: Knowledge Bases, Taxonomic Change, and Contentions with Consensus

> *There are … a number of distinct ways in which a power may vary in the directions of increase or decrease…An obvious dimension of power is that of* ***extent****: if one's control extended over the entire bibliographical universe, the extent of control would be as great as possible. The actual extent of a man's power would be specified, if this were possible, by enumerating or describing in general terms the items over which he had control, items which I shall say constitute the field of control … A further dimension we can call* ***range****, though perhaps versatility would be more apt. The range of my power over any collection of objects corresponds to the number and variety of demands I can make on these objects. The greater the number and variety of purposes for which I can have suitable textual means provided, and the greater the number and variety of neutral descriptions in terms of which I can successfully pose requests, the greater power over my field of control … Another dimension is what we may call the dimension of* ***supply****. If we think of the sorts of power we might have over any collection of objects whatever, it springs to mind that the highest degree of power would be conferred by absolute ownership. The owner of an object, if his ownership is absolute, can do what he likes with the object; he can destroy it, mutilate it, give it away, use it in any way that he likes and can.*
>
> *—Patrick Wilson*
> *Two Kinds of Power: An Essay On Bibliographic Control* (1968, pp. 36–38)

Introduction

Given the described differences between traditional, *descriptive-oriented* classifications, and pragmatic, *retrieval-oriented* management classifications, one might ask the question, then, what is the ultimate *purpose* of these composite systems as it pertains to the biodiversity ecology? Based on the evidence presented thus far, I might simply state that the purpose of *systems* (instruments) like the Catalogue are to *organize* context-specific data and information (documents) to *provide access to* scientists (*best textual means*), who can then, in turn, produce more *science* (the situationally-relevant "*means to an end*" in Wilson's bibliographic schematic) (Wilson, 1968, p. 22). But as the quote by Patrick Wilson that opens this chapter indicates, the kinds of powers we have over systems—and their limitations—are essential qualities to grapple with if we are to understand the extent of potential control we have over the documents in question. This chapter examines the potentiality and limits of the Catalogue's taxonomic instrument, which we have spent these chapters deconstructing, identifying, and formulating.

In identifying the *purposes* of information organization frameworks, Joseph Tennis states that they are created for "retrieval, attestation, and inference" (2006, p. 305). Retrieval is one of

the more obvious elements in taxonomic spaces and a prevalent topic in our previous discussions of biodiversity taxonomic structures: the ability to find information related to some taxon or taxa via a mediated browsing interface or query. Attestations are, according to Tennis, descriptions of concepts as represented in knowledge organizing structures. In the context of this manuscript, taxon concepts are validated through and constituted by a series of evidentiary claims about *how* and *why* they are intellectually formulated to fit within scientific discourse. Tennis's final purpose of knowledge structures, that of inference, is particularly important in this chapter. Inferential purposes are the ability to "identify particular documents" (2006, p. 306) within the documentary universe. But given the extensive function of the Catalogue it goes beyond the identification of the represented documents *within* the database, it is also the identification of the purposes of the taxonomic document itself. It can also be those actions that intend to use the Catalogue space to draw conclusions ("inference, n.," 2016) about larger biodiversity questions and frameworks that exceed the additive value of the component parts of the taxonomic structure in-and-of itself. The inferential purposes of the Catalogue exceed the sum of its documentary evidence, associated species concepts, and taxonomic relationships. Instead, inference takes these component elements and applies them to deduce more broadly stated heuristic goals about the *development* of biodiversity and taxonomic science, as well as macropattern conclusions about the state of evolution or structural questions about phylogenetic development. As A. Broadfield states,

> But the sciences do not exist only for the purpose of classifying, and therefore to arrange them according to genera and species is to ignore the explanatory connections which dominate the interests of scientists themselves…The several genera are not found in reality arranged in the system in which classification exhibits them; as they actually appear they are always realised (*sic*) in numberless individual instances, separated in time and space, and subject to continual change both in their own conditions and in their relations to one another (1946, pp. 91–92)

Thus, the Catalogue should optimally be able to provide a flexible environment to facilitate the *application* of concepts to questions asked about some set of conditions of the external world—a world that is continually changing and elusive by its very nature. The biodiversity taxonomic infrastructure should provide an equally-mutable combinatory space where trends and conditions for biodiversity science can be assessed by applying the concepts and relationships the Catalogue presents.

An essential question becomes what should scientists ideally be able to *do* with these composite knowledge bases to answer broader macropattern, evolutionary questions? And, secondarily, what criticisms are applied to the Catalogue that precludes it from having adequate *extensive* uses in professional contexts? What are its *extensive* limitations in the current bioinformatics environment? As an infrastructure meant to deeply integrate itself into an iLife consortium that is fundamental to the global production of scientific biodiversity knowledge, the editors, staff, and facilitators of the Catalogue know that there is much work to be done to meet these various institutional, individual, and systemic needs. If anything, the NAMES conference in Leiden served as a venue where these kinds of limitations could be discussed, so that an adequate plan could be articulated going forward to rectify these limitations.

To this end, this chapter is organized into three broad sections: the first, briefest, section will lay out a possible schematic in which we can understand the use of systems such as the Catalogue for the production of evolutionary and phylogenetic *inference*. Such a discussion situates the Catalogue not only as a *data*base, but also as a space in which this data can be properly invoked to answer larger scale historiographic questions about how *taxonomies* (in contradistinction to nomenclature) can be traced as an evolving series of postulations over a broad period of time. The practice of nomenclature, and more specifically nomenclators, are to

provide a general sense of the historical development of species concepts. The Catalogue-as-knowledge-base can also be conceptualized to provide just such a tracing mechanism for the production of taxonomies, varied and continually-changing as they are as a practice that argues a certain position about (current and ancestral) biological relationships. One mechanism invoked in the database space to maintain these historical taxonomic changes is Walter Berendsohn's (1995) concept of *potential taxa*. Potential taxa are, in retrospect, a seemingly simple concept allowing for the conservation of linkages between alternative taxonomic specifications as they evolve over time. But they hold powerful potential in biodiversity database environments.

The second part of the chapter uses Joseph Tennis's (2015) concept of second-order classification theory to understand the kinds of *possible* changes classification structures undergo, as well as what we do with them, once they have been constructed and implemented in a particular context. Such a discussion frames the changes and uses of the Catalogue within a clear Information Studies position. Taxonomic ontogeny is introduced as a way in which we can theoretically understand the evolution of taxonomic *documents* similarly to the way we can understand the development and history of the species document. One example offered of an ontogenic transformation is the iterative and punctuated change the Catalogue undergoes as new GSDs are introduced into its management infrastructure as it grows over time. Each time a new GSD is absorbed by the Catalogue, a new network of classificatory relationships (spatial and temporal) is introduced, which I call a reformulation of *classificatory distance*. The second use of the Catalogue is how it serves to connect and reconcile various taxonomic approaches by virtue of its management hierarchical function. Such a role is central to the Catalogue's purpose of *communicating* biodiversity information rather than arguing a certain evolutionary or phylogenetic position. And finally, the extensive re-use of the Catalogue is formulated in

Tennis's terms, showing how the Catalogue can mediate between local and global taxonomic practices as it is embedded into other online biodiversity infrastructures as a taxonomic backbone.

Finally, in Part III, the core of this chapter, I examine how these second-order issues and concerns are problematized and limited in the online biodiversity taxonomic and data environment. What are the most fundamental critiques of the Catalogue in relation to its intended and stated *purpose* in the iLife consortium? This is certainly not an exhaustive listing of the various critiques of the Catalogue, only those limitations that were identified most often in the interviews I conducted, and in the online discussions and taxonomic literature I closely examined.[87] The first critique is the conflict between the Catalogue's general aim to collocate disparate taxonomic functions in one system within a funding environment that privileges new and multiple cutting edge projects. The end result is that the Catalogue's basic structure is potentially unsupportable in the long-term unless scientists can centralize administrative and funding allocation (processes that are currently being examined for implementation). Secondly, is the ability for users to comprehensively assess the *completeness* and *quality* of the Catalogue's component taxonomic parts. As was discussed in the previous chapter, in most circumstances, GSD taxonomies can be assessed based on individual or group reputations and expertise. In the Catalogue's diverse database space, however, *assessment* of data is a more problematic issue given the many sources from which the data is collected. This fact makes it difficult for some scientists to implement its data in research endeavors with full confidence. Some general reasons for this situation, as well as some methods articulated to possibly offset this issue, are then

[87] The Catalogue is very much a project in process, so these critiques should be seen less as static limitations and more as possible future trajectories and the identification of natural systemic challenges that occur when management taxonomies meet the function and purposes of traditional classificatory approaches. The Catalogue is aware of these issues and much of their efforts are directed to trying to meet the needs of various constituents. No one system will and can meet every need.

introduced. Given the interconnected nature of the iLife consortium, data *error proliferation* is then discussed as an issue that arises in fragmented networked environments such as the iLife ecology. The fragmentary information space these systems comprise also makes it difficult for feedback mechanisms to rectify these identified errors. The final limitation of the Catalogue we will discuss arises, ironically, from the highly curated composition of its taxonomy. Due to the tightly controlled editorial process for the construction of management hierarchy, it makes it difficult to quickly implement and radiate changes into the biodiversity online ecology. The annual publication schedule for the Catalogue means it takes considerable time for corrections to get integrated into the database of record. Further, the Catalogue's current structure precludes the addition of taxon token forms that fall outside of the nomenclature practices controlled by the various Codes. The result of this approach is that un-named genetic barcodes and other OTU-identified taxon labels do not fit into the Catalogue's structure.[88] One possible approach constructed by Global Biodiversity Information Facility (GBIF) is then discussed as a possible remedy to this general issue.

Taxonomic informatics practice is a complicated endeavor; no database is and can be perfect, it is merely a representational model of biological and scientific processes. Trying to accommodate the complex arrangement of multiple taxonomies, as well as document how those arrangements evolve over a period of time is no easy task, particularly when such complexity

[88] As has been alluded to in this manuscript, this is an issue on the radar of the Catalogue and associated iLife platforms, and meetings such as the NAMES in November conference were meant to begin articulating ways to solve this issue going forward. Recall as well that the Catalogue of Life Plus—discussed in chapter three— is also an infrastructure articulated to iteratively manage this issue (See Figure 13).

must somehow be fixed in database fields that are relatively limited in their composition.[89] As A. Broadfield indicated,

> Sciences are more engrossed with deductions than with placing things in a scheme, and with changes, rather than with the unchanging essences whose nature the theory of predictables would have us investigate. From the scientific viewpoint classification is only preliminary; we may posses so little knowledge that we can only classify, but science cannot long rest content with this (1946, p. 93).

The end result is that preliminary data management activities—in the form of the production of management hierarchies—are occasionally at odds with the contextually specific activity of studying evolution, phylogenetic, and ecological concerns. The Catalogue is attempting to provide a flexible platform in order to be *useful* to scientific practitioners, but in the process of providing this flexibility, it must also manage an individual identity while inhabiting a hybrid space of translation and multiplicity. This chapter chronicles the Catalogue's goal of providing a knowledge base standard, and presents the various complications this endeavor exposes as a function of taxonomic practice.

Part I: Toward Combinatory Knowledge: Macropatterns and Historical Taxonomic Concept Repositories

One result of the increased extensive ability of the Catalogue's biodiversity database system is how it can potentially be used as a springboard to assess larger trends in the taxonomic world, over large swaths of time. As we have witnessed thus far, one of the key aspects of articulating taxonomic knowledge in digital spaces is being able to distinguish and *record* the various *transformations* species concepts undergo historically. Nomenclatural practices, designed as they are to trace the historical development of concepts via a variety of code mechanisms, do not have a direct analogue in the domain of taxonomies. That is to say, we cannot yet (easily) track the minute changes *taxonomies* undergo over time. Such information

[89] Refer to the collection management database limitations for the NHML in chapter one for another example of this issue.

could help us understand the broader trends and developments regarding how scientists have, historically, understood the ontological constitution of the natural world. As taxonomist Nico Franz has noted, "the real-life challenge for these information repositories is to capture more than one authoritative classification; they are built to represent the full spatial and temporal dynamic of the taxonomic process" (2005, p. 499). Dynamism is an essential concept here, especially since taxonomy (and the interpretation of the species in the natural world), as a scientific practice, is nothing if not constantly changing. The infrastructures that we build to support and document these processes should be able to mirror the various transmutations of taxonomic opinion and representation.

As I have emphasized thus far in this manuscript, one of the primary ways in which we can understand the *extensive capacities* of structures such as the Catalogue of Life is through the implementation of its taxonomy as the backbone of *other* infrastructures (Species 2000, 2015b). Another way in which we can think about this *extensive capacity* is the extent to which the Catalogue can provide the capabilities to answer (and infer) "macropatterns" (Bourgoin, 2016) given the documentation of historical concepts within the database structure. In Figure 21, Thierry Bourgoin illustrates a possible schematic for how we can think about this phenomenon. Creating the management classification is just the first step of a long-term initiative by the Catalogue. The Catalogue, as it is currently composed, is a snapshot of current biodiversity knowledge, but that snapshot is relatively static (particular if we think of the most authoritative, Annual version of the Catalogue as the citable database of record). Names-as-data are synthesized and recorded so that valid taxa can then be utilized and implemented as the building blocks for other projects, initiatives, and investigations. In Figure 21, Bourgoin makes a distinction between classification, phylogeny, and evolution. The classification is but the first

preliminary step of this process of examining biodiversity (to invoke Broadfield's statement above once again). The ultimate, long-view (Ribes & Finholt, 2009) goal here for platforms such as the Catalogue, is being able to add the "confrontation of classifications with phylogeny" in such as a way that these systems can better "understand evolutionary [concerns]" (Bourgoin, 2016). The Catalogue, in addition to being a repository of current nomenclature, must also think about how to position itself to also be a repository of historical taxonomic concepts.

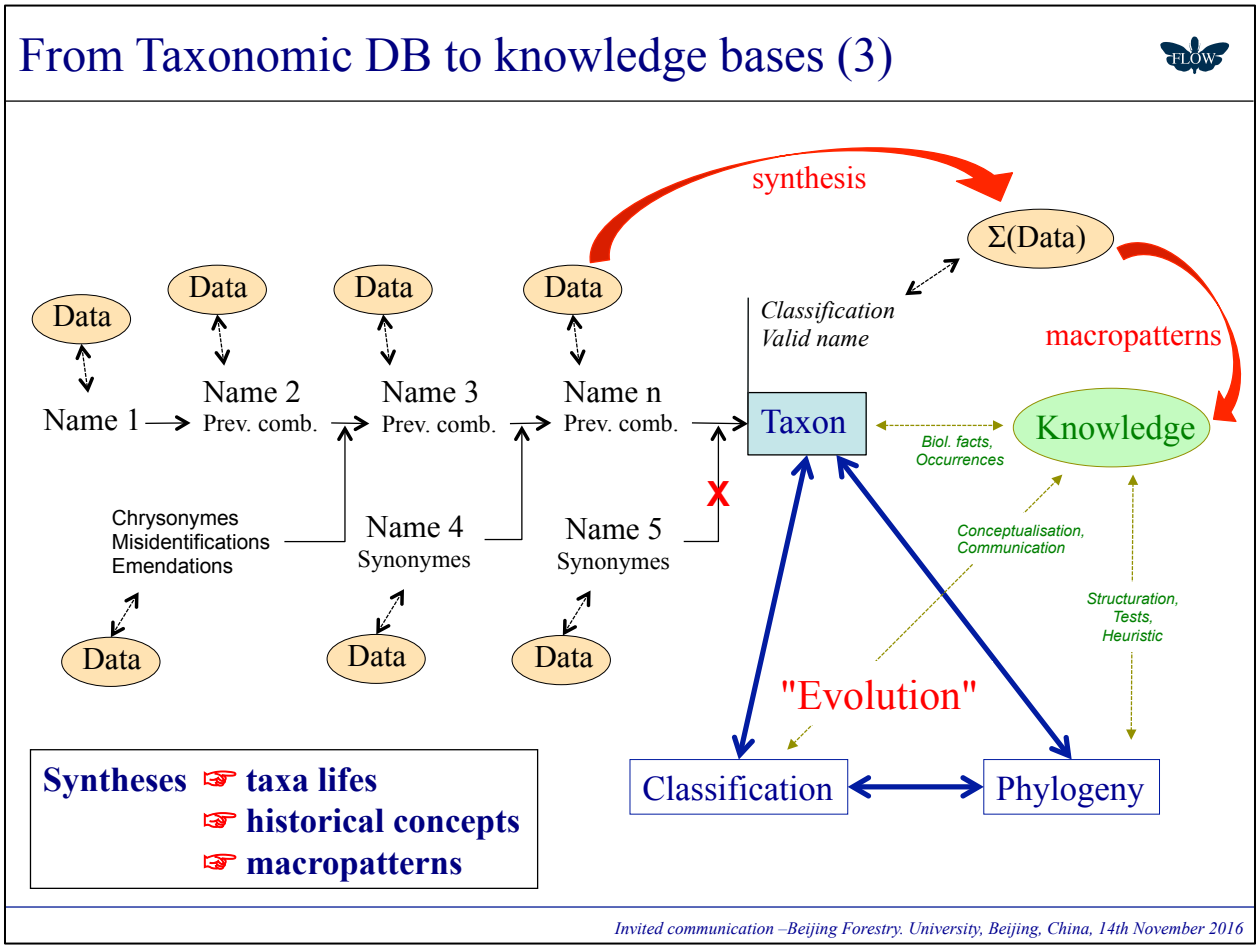


Figure 21. From Taxonomic Databases to Knowledge Bases: Understanding Evolution. This figure expands the view of nomenclatural and classification control seen in Figure 12 to include the assessment of macropattern examinations such as phylogeny and evolution. The goal in a system of this nature is to a) take name data that is synthesized into validated taxon forms, and b) use these taxa as the building blocks for classifications that can then be used to articulate many possible phylogenetic hypothesizations. In a broader sense, the relationship between all of these apparatuses (names, taxa, classification, and phylogeny information) is to understand bird-eye view evolutionary concerns. Such a platform is about using the documentary units of a space like the Catalogue to form a knowledge base and laboratory for speculative evolutionary informatics. Source: Thierry Bourgoin (Bourgoin, 2016).

Biologists and informaticians have termed this kind of macro-level analysis, *evolutionary informatics*, which "concerns the capturing, storing and integrating of all these data (about biological specimens, images, genomes, etc.), as well as developing the analytical techniques that use them to answer evolutionary questions" (Parr et al., 2012, pp. 94–95). Infrastructures such as the Catalogue of Life stand at the center (Parr et al., 2012, p. 100) of this informatics integration, as either the taxonomic backbone to platforms like GBIF, or as the aggregator of curated taxonomic information and biodiversity data from WoRMS, ITIS, and any number of other infrastructures. This makes the Catalogue a key player in the constitution of any knowledge base, framing species occurrence data into composite taxonomies structures that can be recombined in numerous ways. The sum of this entire ecology of historical documentation provides a sandbox of *potential* knowledge products beyond the database content itself. Such work recombines data in ways that can provide long terms predictions, climate and species modeling and simulation (Landers, 2016).

The success of this kind of data integration, however, involves the careful documentation of how taxon groups *change* over time. It is generally acknowledged that the long-term prospects and success of the Catalogue, in one sense, depend upon the system's transformation from one that organizes taxa to facilitate data sharing, to a system that can map the historical contours of the various taxonomic instruments it inherits to create its superstructure. To do this, the original taxonomies and taxon concepts for GSD databases need to somehow be maintained as they intermingle in the management hierarchy. As we have seen, the core building blocs of this activity lies in an as-comprehensive-as-possible pool of species names (as concepts), so the Catalogue of Life must continue working to bolster the quality of its nomenclature foundation to assure that there are units to recombine and translate between different taxonomic specifications,

approaches, and methodologies. The Catalogue of Life Plus (discussed in chapter three) proposes a mechanism through which the historical trajectories of nomenclature can be documented. However, as we have seen, "due to the inherent limitations of nomenclature a name may correctly designate several perhaps equally well-founded concepts of a taxon" (Berendsohn, 1995, p. 210). The question then becomes how to record and maintain these multiple interpretive name forms in the Catalogue database *after* they get rearticulated into accepted taxa and embedded into the management hierarchy? Once this kind of activity is perfected, it can serve as a kind of historical taxonomic concepts repository.[90] To answer this question, we briefly look to the concept of *potential taxa*.

In 1995 Walter Berendsohn published an influential article titled, "The Concept of "Potential Taxa" in Databases," which provides a possible way forward for documenting historical taxa interpretations in digital spaces. As we know, names stand for species concepts, but those concepts change over time. Nomenclature forms provide some mechanisms for tracking these changes, but the challenge becomes how to interconnect these variances within the spaces of databases and database structures.[91] A "potential taxa," is one solution to this problem, which "is a name with taxon circumscription information attached to it by means of one or more literature references" (Berendsohn, 1995, p. 207) that is then connected to other name forms within the database environment using relational fields. Continuing, Berendsohn indicates how such a concept can be implemented within the biodiversity database structure itself,

> Due to the inherent limitations of nomenclature a name may correctly designate several perhaps equally well-founded concepts of a taxon. For the purpose of information handling, a way has to be found to differentiate between different taxa bearing the same name. In an information system, this can be achieved by introducing a data element or data area which impartially mirrors alternative taxonomies, and allows for

[90] A historical taxonomic concept repository is one of the potential long-term goals articulated at the NAMES in November meeting for a species list such as the Catalogue of Life.

[91] See my example in chapter three, adapted from Richard Pyle (2008), on the importance of nomenclature forms to trace historical changes to species concepts in a hypothetical fish population.

> the inclusion of all information-bearing individual taxonomic concepts, including misnomers….A database system using potential taxa is able to treat an unrestricted number of different concepts related to a specified name. In a relational database system, this does not pose a technical problem, because the entity type "potential taxon name" must have but three attributes: a pointer to a name, a second one to a circumscription and status intersection which handles the name status and the connection to the circumscription reference, and finally a pointer to an entity-type handling classification. (1995, p. 210).

As databases enter the Catalogue database, the 'holding space' field of potential taxa allows for mapping and re-mapping of taxon concepts over time. Additionally, potential taxa relationships allow for the ingest of taxon concepts into management hierarchies without losing their original taxonomic context and circumscriptive context. Names can be mapped to any of their previous taxonomic instantiations depending on how one chose to orient a particular query. Such a network of taxon connections records all of the possible name and data combinations depicted on the left side of Bourgoin's Figure 21. With this recorded historical knowledge, new classificatory and phylogenetic arrangements can be proposed using this robust metadata. As H. Charles J. Godfray (2002) has indicated, the goal of a unitary taxonomy, as represented by a space such as the Catalogue of Life, is that it can serve as a single reference structure for "accumulated knowledge," and that such a repository is essential in an age of information fragmentation. However, a downside of such an approach—especially one that does not take potential taxa into account—is the erasure of taxonomic expressions (2002) in these in composite structures. The concept of potential taxa allows unitary taxonomic approaches to "easily provide information on concept relationships between different systems and treatments thus creating a pathway between current and past treatments" (Berendsohn & Geoffrey, 2007, p. 20).

However, Berendsohn identifies a few downsides to adding potential taxa metadata in biodiversity databases, including the inflation of species records, as well as a potential deluge of information for users in retrieval searches (1995, p. 211). In order for databases to function as a kind of knowledge base proposed by Bourgoin, however, such issues must be accepted and dealt

with procedurally as part of the database interface. Additionally, as platforms like the Catalogue of Life increase their role in linked-system environments, such as the iLife ecology, maintaining taxon concept changes over time is going to be especially important. The dynamic structure of the Catalogue can then be documented and traced using the ever-increasing network of potential taxa connections. Potential taxa will undoubtedly continue to play a major role in the extensive power of systems like the Catalogue as a knowledge base for generative research. One must be able to identify the kinds of changes taxon concepts have undergone and juxtapose this knowledge with current scientific knowledge.

Moving onward, building on this notion of taxonomic *change*, in Part II of this chapter, I will invoke Joseph Tennis's concept of *secondary classification theory* to examine three kinds of transformations that taxonomies like the Catalogue must contend with as part of their construction and use. We can then proceed to some of the specific critiques of the Catalogue that stem from some of these transformations.

Part II: Taxonomic Change, Interoperability, and Transformation

One of the most prominent issues encountered in biodiversity database systems is the extent to which GSDs are "qualitatively transformed" (D. Remsen, 2010) as they are ingested into the compiled system. As databases enter the Catalogue of Life system, the editors do their best to maintain the integrity of the original GSD structure, but there are times when "adjustments may need to be decided upon by the editors on where and how to insert it, to make it as consistent as possible [with the rest of the Catalogue of Life], while not losing the essential taxonomic information it has been created to provide" (Species 2000, 2016c). One need look no further than Figure 19 in the previous chapter, which compares the taxonomic hierarchy from both the "Species 2000 & ITIS Catalogue of Life" (the upper image) and the standalone

"Integrated Taxonomic Information System (lower image)," to see how this kind of editing takes shape within database environments. The ITIS taxonomic hierarchy is far more detailed in its composition, particularly because it contains a more finely articulated higher-level taxonomic ranking system (inclusive of subkingdom, subphylum, infraphylum, superorder, etc.). In comparison, looking at the Catalogue of Life's taxonomic tree, you can see that the Catalogue editors chose the *genus* node as the connection point for the ITIS database. Note the lack of an author and publication designation after the species epithet, *Ursus*, at the genus level in the Species 2000 taxonomy, which indicates that *Ursus* is part of the less-detailed Catalogue of Life management hierarchy. The Species 2000/Catalogue of Life hierarchy, on the other hand, retains the ITIS taxa detail below the genus node connection point (beginning at *Ursus arctos* Linnaeus 1758), while above that point the general management classification backbone connects the ITIS classification to the rest of the contributed GSD database sets. The *act* of choosing one part of the ITIS classification, over any other, transforms the ITIS structure in fundamental ways, divorcing *Ursus arctos* from the upper backbone context in its original coordinated location in the ITIS database. Keeping this example in our mind as a concrete model, our attention will now turn to the kinds of ontological changes and taxonomic transformations that take place by virtue of the Catalogue of Life's approach to taxonomic management. To do this, we turn to Joseph Tennis's metatheoretical framework for classification structures to help us frame and define the kinds of *change* these taxonomies are engaging in over a given period of space and time.

Contours of classification.

In assessing the metatheoretical "contours" (2015, p. 244) of classification theory literature, Joseph Tennis has identified three basic strata, or approaches, to the study of classification that define the kinds of work currently being performed in this domain. Tennis's

categories include, foundational classification theory, first-order classification theory, and second-order classification theory. As defined by Tennis, foundational classification theory is "concerned with philosophical and definitional aspects of classification" (2015, p. 246). A. Broadfield's (1946) work, which has been invoked numerous times in this manuscript, is an example of such an approach, as would much of Bliss's categorical and philosophical expositions (Bliss, 1929, 1933)[92]. Larger questions in this arena include ontological and epistemological questions about the concept and act of organizing knowledge. First-order classification is "solely concerned with the methods of classification scheme construction and use" (2015, p. 245)—very straight-forwardly: how we *build* classifications and articulate the processes we create in order to produce them. The two previous chapters pivoted on such first-order approaches: how the Catalogue allocates evidence for concepts, how such concepts can be connected within the nomenclatural system, and how management classifications build their schematics in contradistinction to traditional taxonomic forms.

Finally, and most pertinent to our current discussion, is second-order classification theory, which is "concerned with what to do with classification schemes once they are built" (2015, p. 246). In the case of the Catalogue, we are concerned with how GSD taxonomies are *changed* and manipulated in order to be absorbed into the Catalogue's taxonomic space. Secondarily, we are interested in how its management structure is used in various contexts and how that intended use fundamentally changes the Catalogue's composition. In practice, of course, any type of work performed in the domain of knowledge organization is likely to function on a number of Tennis's levels simultaneously, since, for example, one cannot examine how to define a classification system without addressing the way it is *constructed* and then

[92] Bliss's texts have a tendency to include *both* foundational and first-order discussion, the latter often introduced after broad epistemological and ontological questions have been adequately discussed as a contextual ground for his analysis.

subsequently *used*. But for our purposes we will focus more on the general notion of taxonomic *change*, both of the Catalogue's taxonomy itself as well as with the taxonomies that comprise the Catalogue's backbone. These concerns relate to how a composite taxonomy specifically is essentially used, and how such use requires change under certain conditions. These issues stem from the *composite* nature of the Catalogue, and fall under Tennis's second-order classification, which is broken down into three subcategories, as follows: (1) how schemes change over time and how we update them, (2) how installed schemes interoperate, and (3) how systems change when they change context (reapplied or reengineered)" (2015, p. 246). While each of these elements are pertinent to biodiversity databases, the second and third are most relevant for the case of the Catalogue. Keying-in on Tennis's second-order classification issues, we will discuss these in turn as they relate to the Catalogue.

Taxonomic scheme change: Ontogeny and the taxonomic document.

Taxonomic schemes change, particularly as the conception of the classified biological *object* (species concept) are redefined over a period of time. Such ontogenic concerns (Tennis, 2002, 2012, 2015) have been of increasing interest in the Information Studies community, concerned as we are with how to theorize and manage the lateral transformations of knowledge organizing systems to keep them updated and relevant to our organizational and access concerns. As Tennis notes, ontogeny is the examination of "the life of a subject over time—the subject's scheme history" (2012, p. 1351).

Ontogeny, or ontogenesis, is a surprisingly appropriate term for the development of classifications in our context, given its general usage in the field of biology to describe the mechanism and development of individual organisms. In Ernst Haeckel's, *Generelle*

Morphologie der Organismen (1866), he defines four general concepts: ontologie, phylogenie, ontogenesis, phylogenese. Løvtrup summarizes Haeckel's concepts in the following way:

> «Ontogenie»: the history of the development of the individuals.
> «Phylogenie»: the history of the development (evolution) of the taxa.
> «Ontogenese»: the mechanism of the development of the individuals.
> «Phylogenese»: the mechanism of the development (evolution) of the taxa (1987, p. 199)

In practice, as Løvtrup has indicated, the distinctions between the *mechanism* and the *history* of both ontogenetic and phylogenetic processes are no longer upheld (1987, p. 199), and as such the two Haeckelian term-pairs are now synonymous in discourse. Ontogenie (and its partner term ontogenese) refers to individual organisms—as in the progressive and regressive (1987, p. 201) development of one organism over its life span from its embryonic stage to death. [93] Phylogenie (and Phylogenese), on the other hand, refers to the historical reconstruction of taxon groups that, at base, are represented in phylogenetic classifications. But how can such terms be thought of analogously to the way we think about *classifications* in the biodiversity world? For one, and perhaps the most obvious, is the ways in which classifications are, or at least can be, a representation of an expressed and argued phylogenetic relationship. A given phylogenetic classification, as Løvtrup indicates, is one or "many [possible] lineages of progressive evolution" (1987, p. 201). The historical reconstruction of *individual* development (more intensional in nature) is "currently covered by the term <<developmental biology >> or << epigenesis>>" (1987, p. 199), and is generally similar to Tennis's ontogenic examination of subject terms over time.

Our more documentary-focused concerns are about how species concepts are represented by name tokens, as well as how those tokens are coordinated in an overall taxonomic framework reflected in certain representational database structures. If we think about the change of subjects

[93] Løvtrup defines the progressive phase as the "phase, lasting usually from fertilization to maturation," and the regressive phase "from then on until death" (1987, p. 201).

over time (in the sense that Tennis invokes above), the subjects of biological classifications are species (or species concepts articulated using a number of subsidiary forms of evidence). And, as we have seen, the *position* of that particular subject in a given classification is dependent upon the relationships built within that *instrument*, conceived as a cohesive whole based on the classificatory and methodological commitments of its builder. The *instrument* itself (and the Catalogue, in particular), then, is a document that can also itself go through various ontogenic transformations—defined by the accumulation of subject changes over time, as well as the changing positionality of its 'subjects' within a constellation of related concepts.

Such change happens on numerous levels in the Catalogue, one being the extent to which GSD taxonomies are snipped out of their original context for inclusion into the Catalogue of Life.[94] The implementation of proto-GSDs in the previous chapter is another good example of such change: taxonomies are combined in various ways to make functional taxonomic systems to fill taxon gaps. A key problem in the Catalogue's space is the extent to which the changes imposed onto GSD systems manifest in greater or lesser collocative capabilities within the Catalogue (Tennis, 2014, 2015, p. 246). The Catalogue's data structure (Figure 22) is articulated in such a way to best trace and interconnect any given species concept (name) to its original position within the GSD (or RSD) "source database" data fields (in theory, these linkages are not only internal to the Catalogue, but also external in that they link back to the original GSD

[94] While I have categorized the change GSDs experience as they travel into the Catalogue in terms of Tennis's first second-order concern—how "scheme changes over time"—I acknowledge that this is also very much in accord with Tennis's third kind of second-order classification concern—that of "how systems change when they change context." The assumption here is that many of the changes and uses of taxonomies can potentially fall under a number of Tennis's categories, depending on what element of that change you emphasize. Secondly, I have chosen in this manuscript to focus, not on the changes these GSD taxonomies in-and-of themselves undergo, but on the Catalogue of Life's context in particular. Since the Catalogue of Life is the emphasis here, I see the inclusion of these taxonomies changing the *Catalogue's* scheme structure, which is the *primary context* of note. The GSDs are changing context, not the Catalogue, which is the context. If my emphasis were on the GSD space in particular, then I would have categorized this kind of classificatory change under Tennis's third aspect.

website, if one exists), as well as with the common names, synonym, and GUID's that interconnect it with other platforms in the iLife consortium and beyond.

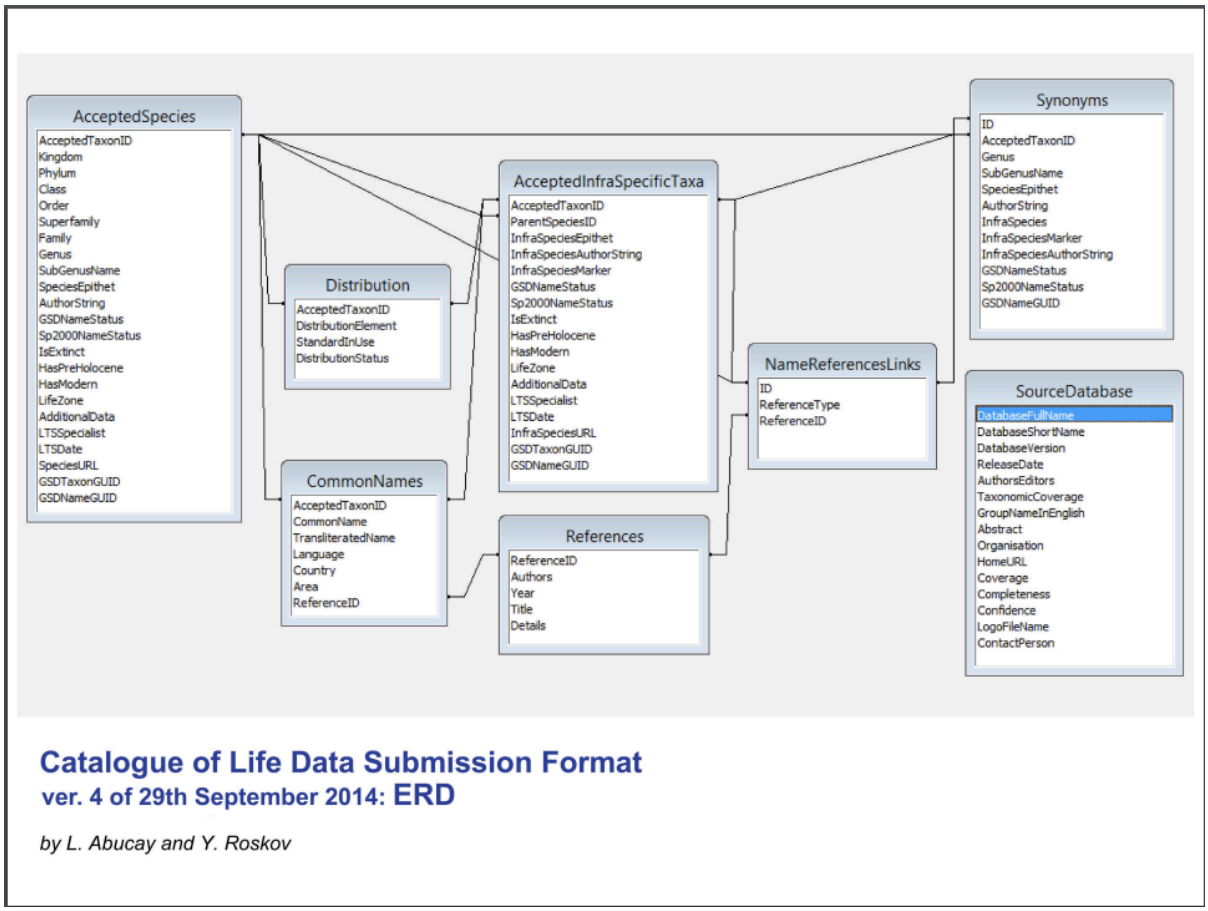


Figure 22. Catalogue of Life Entity Relationships Model, Version 4, 29th September 2014, by L. Abucay and Y. Roskov (Species 2000, 2016f).

The position of subjects (species names) is also continually redefined in the Catalogue's space. As taxonomies are brought into the Catalogue, they are set alongside other taxonomic structures, and by virtue of these juxtapositions the relative positions of one subject in relation to another within the Catalogue's classificatory structure are transformed. These new arrangements have significant effects in terms of how we understand *relationships*. Relationships that are essential to understanding the internal theoretical logic of any given taxonomy. Relationships, however, are not just intellectual associations between classes (Green, 2008), but also spatial and temporal ones. I call these spatial and temporal relationships *classificatory* (or, to use Tennis's

vocabulary, *ontological*) *distances*. In his book section, "Confusion Confounded," Bliss (1933, p. 218) speaks about the placement of disciplines in the Decimal Classification schedule and how the placement of topics has the potential to cause confusion: "Methodology (112) stands between Ontology and Cosmology and far from Logic, of which it is usually regarded as an extension" (p. 218). Olson also points to the definitional qualities of classificatory structures: "the hierarchy thus created structured knowledge by putting every subject in its place. It creates a context for each subject within this hierarchical arrangement" (2002, p. 22). Locations within class schedules and facets, indeed, exist in a kind of Euclidian space with quasi-quantifiable, and certainly qualitative, distances set between them. As Rebecca Green and Giles Martin have indicated, "Traditional (rank-based") biological taxonomies are organized hierarchically, with the rank (e.g., kingdom, phylum, class, order, family, genus, species) of a taxonomic unit/taxon indicating its relative position in the taxonomy" (2013, p. 10). *Relative* is a key term here, since if a portion of a taxonomy is removed from any given GSD and inserted into the Catalogue's taxonomy, its definition within a particular hierarchy *relative* to other subjects has changed dramatically within this new context.

This definitional problem is compounded when two *distinct* species share the same name within the Catalogue due the existence of different nomenclatural codes governing different parts of the biological kingdom. Recall that codes dictate the application of names (ICZN and ICNAFP, for example), and that each code prohibits the duplication of names for multiple species (homonymy). However, since the Codes are independent of each other, it is possible to have two species with the exact same names in both the plants and animal domains (Global Names Architecture, 2016; T. Orrell, personal communication June 15, 2016). As the International Commission on Zoological Nomenclature states, "for example, the genus Ficus is

available and valid for both a gastropod genus and the plants commonly called figs. It is assumed that points of confusion in referring to organisms in different Kingdoms will be rare, thus homonymy is not controlled in these cases" (2017). In the Catalogue of Life, which is ostensibly charged with organizing multiple names following rules under different codes, such homonymy, though rare, happens on occasional circumstances. The existence of these possible errors also changes the definitional status of certain subjects given that they can no longer be disambiguated as they had been in their original GSD context.

Lastly, the *inherited* qualities of *descriptive-oriented* classifications also include temporal associations that are transformed in the Catalogue's new taxonomic space. "As a temporal mélange, the [Catalogue] is not an internally consistent system" (Montoya & Erickson, forthcoming, p. 3). The evolutionary, cladistic, and phenetic schools—and the representational diagrams they produce—all have within them assumptions about not only an organism's formal and physical relationships, but also an implicit statement about how they are related in an interpreted evolutionary, genealogical, or ancestral framework. In evolutionary taxonomies, for example, classificatory distance "[depicts] genealogical history over time " (Podani, 2013, p. 322), and thus a shorter or longer classificatory distance distort the temporal relationships between subjects. Again, while management classification of the Catalogue is "hierarchical and reflective of phylogeny," it is not "itself a phylogenetic tree" (Ruggiero et al., 2015a, p. 3) due to its practical approach of intermingling taxonomic approaches, so temporality is less of an issue in this composite space.

The schematic transformations exhibited (and elicited) by the Catalogue apply to two distinct "subjects" within the taxonomic world: the species concept subject, as well as the taxonomic document subject. Each of these entities undergoes various changes as GSDs get

embedded within the taxonomic space of the Catalogue of Life. Hand-in-hand with how these subjects change is how the Catalogue is designed to operate, *extensively*, with the various other iLife infrastructure—the subject of our next section.

Classificatory interoperability and reconciliation.

The Catalogue of Life data submission format (Figure 22, above) is also pertinent to Tennis's second second-order classificatory concern: that of how installed schemes interoperate. The data structures implemented by the Catalogue of Life facilitate ready transfer of data between multiple biodiversity communities.[95] Indeed, the interoperability and evolution of taxonomic infrastructures has been recognized as a vital subfield of inquiry in I/S and KO, particularly given the rising need for computational infrastructures to communicate with each other effectively in a networked environment (Jung, 2008; Lei Zeng & Mai Chan, 2004; Teckelmann, Reich, & Sulistio, 2011). A 2015 one-day conference in Copenhagen, titled "Global and Local Knowledge Organization," was meant to promote "a conversation about the tension between the global information structures and grounding meaning and ethics in localized contexts" (Mai, 2015)[96], and further illustrates the growing interest in taxonomic interoperation. Laura Skouvig's presentation at this conference expressed the context-specific nature of 'information,' and how notions of 'global' information are often in discord with local knowledge systems (2015). Skouvig outlines how the control of information by the media in Denmark (the 'global') has historically worked both harmoniously and in conflict with local conceptions of

[95] Darwin Core (Darwin Core Task Group, 2011) is the most generally accepted data standard and one of the possible submission formats for the Catalogue of Life (Roskov, 2016c). Darwin Core, however, does not have attributes or concepts that facilitate say, the description of host associations, which are valuable pieces of ecological information: "[Host associations are] not part of the standard Darwin Core. If we have this species A and it's on Plant B, Darwin Core doesn't have that set up properly to convey that information" (Yanega, 2016). Local repositories, then, often create idiosyncratic standards, on top of Darwin Core, that can facilitate this kind of species/ecological documentation.

[96] See also (Adler et al., 2016).

information (as in exerting absolutist power by controlling the dissemination of information); in order to understand this reciprocal relationship, Skouvig claims, one needs to understand the historical and geographic context in which these practices evolved.

Similarly, composite taxonomic authorities such as the Catalogue of Life must be understood within a historical, intellectual, and geographic framework in order to understand how international (global) standardization works in concert with, and against, the local repositories (individual biodiversity scientists or teams) that contribute to them. At its heart, the Catalogue is nothing if not a platform designed for the "switching and reconciling" of different biodiversity relationships, "vocabularies[,] and by extension, classification schemes" (Tennis, 2015, p. 246). Previous chapters in this manuscript have explained how nomenclatural practices and codes have been put in place to facilitate the ready communication of species concepts, as well as to maintain links to the evidence used to support these circumscriptions. As an aggregative system, the Catalogue's management hierarchy is meant to help taxonomies communicate, knowing full well that while it is not a perfect solution, a global view of biodiversity knowledge is necessary to understand the state-of-the-discipline. Such reconciliation is essential, especially if other iLife infrastructures, such as GBIF, can append occurrence records in the form of data points that "document evidence of a named organism in nature" (GBIF, 2016a). The extent to which the Catalogue can serve as an effective "switching mechanism," however, is dependent upon the size of the system that integrates it, the purposes it serves, and the breadth of data it concerns itself with, which we will examine later in this chapter in relation to GBIF's use of the Catalogue's taxonomy.

Taxonomic transformations: Re-purposeability and extension

Tennis's final second-order concern has to do with how classifications change when they are *transformed* (2015, p. 246). How the Catalogue is repurposed and *reengineered* is absolutely critical if we are to understand the limits of the Catalogue's *extensive* capabilities. Tennis articulates that, in "cases [where classifications are transformed from one kind of structure into another] we capitalize on the loose definition of classification in that we see concepts and relationships that obtain between relationships and we feel open to modify structure, often adding functionality, but sometimes taking it away" (2015, p. 246). As I claimed in the previous chapter, expanding Wilson's notion of bibliographic power included the coordinated, controlled, and open design of the Catalogue for its re-purposing and use in subsequent systems. Each of the systems in the iLife consortium in Figure 1 implements the Catalogue's taxonomy to some extent or another. In the case of the Encyclopedia of Life, the taxonomy is displayed as one option among an array of taxonomic approaches; in the case of the GBIF, their taxonomic backbone builds upon the Catalogue's hierarchy to organize the data compiled from sources that may not otherwise have this intellectual structuring agent. One might think of this as a *taxonomic amplification*—the use of a composite taxonomy to point to GSDs that hold more in-depth caches of primary data; collocate multiple sources of taxonomic information; and, finally, to produce more taxonomic knowledge by the recombination of taxonomic information.[97] The Catalogue is intended to support the shift from local practices (idiosyncratic, internally consistent) to global spaces and standardization, intended for full-scale integration into the global infrastructure to support the "possibility of a quantum increase in the coherence of the world's biodiversity data and analyses" (Species 2000, 2015b).

[97] Recalling here Thomas Orrell and Gerald Guala's concept of *synonymic amplification* which described the process of making biodiversity information more locatable with the usage of synonyms deeply embedded into information retrieval systems (2016).

Given these various mechanisms by which the Catalogue is used and changed in the iLife ecology, the next task is to examine how these ideal second-order notions are challenged by its use in extensional practical circumstances. While the Catalogue is certainly a large player in the biodiversity community, it has not been adopted and embraced full-scale by all taxonomists and bioinformaticians. Stephen Thorpe's general critique of the management classification approach can set the stage for our next discussion of the Catalogue's limitations:

> Biological classification is a mixture of scientific fact …and subjective opinion … Both these factors taken together doesn't make life very easy, and it is all in perpetual flux…However, I don't think that the issue can be "managed" in quite the way that is envisaged by some. I have thought a great deal about this, for my Wikispecies work. My primary governing principle is that, subject to monophyly, classification is primarily a filing system to make information management easier. So, it doesn't really matter which classification is followed, PROVIDING that it is explicitly stated which one. The problem with adopting a particular classification for a large group (like the "Protista") is that advances in taxonomy happen on much smaller subgroups, so if you blindly follow one particular broad classification, then you cannot accommodate the advances very easily. Hence, I think you have to simply treat matters on a case by case basis, and just choose and specify a sensible classification for that particular case (and change it, if necessary, if something more convincing is published). To try to come up with a single "officially endorsed" classification would simply be to ignore the subjectivity and fallibility of taxonomy... (2009).

Part III: Limitations of Aggregated Taxonomic Knowledge

The prevailing assumption I have been making about the use of management classifications to organize global species data is that they are an unequivocally useful and necessary instrument for data facilitation. The benefits of systems like the Catalogue are particularly popular to those heavily involved in biodiversity informatics where informaticians understand the need for consistent information control standards for the ready communication of disparate data points—but even in this space, agreement is not universal, as Stephen Thorpe's quote reveals. As Brian Buchanon indicates in *Theory of Library Classification*, "we have summed in this book that systematic order achieved through the use of classification scheme is helpful; but we ought to consider its limitations and the objections to it" (1979, p. 119). It makes good sense to ask the same questions of composite management taxonomies as they relate to the practice of biodiversity work. The second-order issues, as they relate to classification systems

like the Catalogue of Life, lie at the core of what motivates the Catalogue's chosen structural approach: they want flexible structures that keep a modicum of consistency in a landscape defined by change. But this very taxonomic change defines why many taxonomists and practitioners of biodiversity work deny the usefulness and pertinence of the Catalogue to their daily research. As Thorpe's comment above brings to the fore, taxonomic *contingency* will perpetually be in tension with the urge to control that flux in "officially endorsed" systems designed for *stability*. Thorpe, however, does admit to the need for *some* control, adding that "a sensible classification" is prudent in some, practical cases. Yet his statement seems to apply only to smaller taxon groups (the "case") than to the whole 'tree of life', as the Catalogue is attempting to accomplish. What Thorpe seems categorically against is a fully coordinated, top-down approach to classification. Maintaining numerous smaller taxonomies, however, does not facilitate the integrated 'global view' of biodiversity knowledge that platforms like the Catalogue, GBIF, and EoL, are attempting to produce. Again, descriptive-oriented and retrieval-oriented taxonomic systems take different epistemological positions about the scope and method of biodiversity knowledge management; and these two approaches often find no middle point in practice.

Thorpe's view certainly seems reasonable, especially given the fact that such top-down approaches (H. Charles J. Godfray, 2002) to classification are antithetical to the way in which taxonomy has functioned for hundreds of years. Contemporary biodiversity practice, however, is seeing the production of data at increasing rates, and large-scale questions about environmental issues and mass extinctions are driving the need for integrated approaches (Guralnick & Hill, 2009). Fragmented approaches to data curatorship and storage just cannot support current big questions in global research. Additionally, as a prominent biologist indicated to me, taxonomic

"perspective is costly" for advanced users, so maintaining numerous independent taxonomic systems is not necessarily a pragmatic approach. On the other hand, while "the world checklist system [approach works at the] production level, for advanced uses, it's shit," as one prominent taxonomist poetically proclaimed. There are many reasons why scientists draw this conclusion, included aggregated spaces like the Catalogue often reassemble taxonomic structures that specialists require in their original form; it is often difficult to assess data quality in these spaces; taxonomies also often contain errors that are difficult to mediate and fix, etc.

Finding the balance between generalized and expert system usage is easier said (and planned) than performed in practice. As a case on point, in response to an inquiry regarding the level of impact composite taxonomies like the Catalogue have for *professional use*, Dr. Douglas Yanega, Senior Museum Scientist for the University of California Riverside Entomology Museum, indicated that,

> The answer is not a simple one, because the problem of creating a single, accurate catalogue is so incredibly difficult. I know several of the people involved in Catalogue of Life, and both Species 2000 and ITIS, and the ambition is great but the funding (and therefore the reality) are still inadequate to match that ambition. In a nutshell the problem is this: when a catalogue is incomplete or inaccurate, how can you tell what's missing, or wrong? If you don't know, then how much trust can you put in that catalogue? … The bottom line: in order for the Catalogue of Life model to ever be realized, we would need to have every taxonomist in the world supported from a SINGLE source of funding, and submitting their work to a SINGLE authoritative body of reviewers. It's a far, far cry from what we have now, and while I can have the dream, that's all it's every going to be - a dream. I admire the goal, I admire the ambition, and I wish it could work, but I don't believe it ever will. If you were to speak with anyone in the taxonomic community who is not personally involved in the effort, I think you will find virtually all of them treat ANY "top-down" initiative with skepticism, because the top is never a big enough umbrella to include everyone (2016).

Aside from the epistemological issues previously discussed that make a single system nearly impossible to implement on a universal level, two other critiques of importance arise in this context: first, funding models do not support the distributed work of the unified systems, and secondly, being able to judge the efficacy and completeness of the data sets included in said systems. Let us now switch our attention to examining these two issues in turn.

Distributed systems, distributed funding.

As to funding, so much of taxonomic science—like any science, really—is dependent on the funding models that support the desk- and institutional-level work of biologists. The sentiment shared by Yanega regarding funding is a fundamental structural concern expressed by countless individuals I interviewed, both formally and informally. Funding agents just are not interested in supporting the basic infrastructure that makes centralized nomenclature and taxonomic services *work*. Despite this hesitation, institutions have seen much operational improvement by centralizing core biodiversity services. Large institutions like the Royal Botanical Gardens, Kew, and the Natural History Museum, London, have the institutionally backed capabilities, and long-term strategic plans, to aggregate biodiversity informatics and spatial analysis work. Kew's Science Strategy for 2015-2020 (2017d) highlights the importance of curating "data-rich evidence," and the subsequent dissemination of that data in online systems like Index Fungorum (2017c) and IPNI (2017e). Similarly, the Integrated Taxonomic Information Service (ITIS) at the Smithsonian NHMN, has received "a prestigious national award for successfully completing a major project aimed at providing easy access to the first credible database of scientific names of organisms in North America and its adjacent waters" (Office of the Secratary, Catherine Hawcker, 1998). The HAMMER award, created to acknowledge improvement of "[government] service to the American people," noted ITIS's centralization of nomenclatural vocabulary so central to the natural history data produced for six government agencies, including "the U.S. Geological Survey, the EPA, the National Oceanic and Atmospheric Administration (including the National Marine Fisheries Service and the National Oceanographic Data Center), the Natural Resources Conservation Service, the Agricultural Research Service and the Smithsonian Institution's National Museum of Natural History" (1998).

Yet despite this proof that the centralization of local and site-specific data stores brings economic and operational benefits (and, perhaps more importantly, facilitates cooperative and aggregative scientific activity), scaling-up this collaboration outside of specific institutions, and convincing funders of its benefit, is proving especially difficult.

The implications of a centralized CERN-style model introduced in the first chapter of this manuscript are geared toward the funneling of such resources into centralized buckets of activity that support global work. But a centralized database model is less-than-glamorous in the eyes of funding sources, often instead preferring software and biodiversity systems that represent new developments and approaches to information management. These "production level" tools are "very difficult to fund," Alan Paton of Kew conveyed, "people don't want to fund the basic infrastructure that makes [taxonomic research] work. Funding is often available only for the development of new "systems, ideas, and things" (personal communication, August 24, 2016). Science in general, is "all about change," another taxonomist remarked, "and functions within a paradigm that prioritizes research impact" as a mechanism to assess both quality and continued funding. But touting the importance of research infrastructure is a particularly "hard sell" (Borgman, 2015, p. 286) for funding agencies. As Chris Lyle, a Researcher in the Etymology unit at the Natural History Museum, London, indicted, this nomenclatural, taxonomic, and biodiversity work is performed "against a social / political / funding background that encourages intellectual academic research but not basic population of databases and moving from demonstration models to production systems" (personal communication, July 21, 2016).

Funding models, therefore, hamper the process of *maintaining* taxonomic infrastructures such as the Catalogue of Life. Acknowledging that funding agents influence the direction of scientific activity is hardly a new observation. Similar issues arise in terms of biodiversity

research emphases, as Geoffrey Bowker makes clear, with more funding going to charismatic species over species perhaps less appealing, or of direct concern in the medical or vector community (2008, Chapter 4). As successful as the Catalogue has been, it exists on a plane with other platforms that, in theory, are also competing for the same, or at least similar, resources. Over the Catalogue's seventeen-year existence, operations have run on a series of European Commission grants totaling some eight-and-a-half million Euros (Y. Roskov, personal communication, January 20, 2016). Strategic gatherings such as the NAMES in November meeting I attended in Leiden (Global Names Architecture, 2016b), are aimed, primarily, at the coordination of nomenclature, not *only* because it makes good intellectual and professional sense (though, of course, that alone would warrant the meeting), but also so that clear technical *roles* could be articulated to each iLife member. I use the term *technical* purposefully in this context, since the vast majority of grant funds awarded to the Catalogue (and, as I found, to other iLife systems) "go toward [information technology] development and building infrastructure for the Catalogue of Life. Not for the content. It is very difficult to find money for [content]" (Roskov, 2016a). Activities such as proto-GSD creation and gap analysis are incredibly resource intensive, and are added on top of an already heavy taxonomic editorial load to make the management hierarchy function form edition to edition. The vast majority of the Catalogue's global team and taxonomic advisors contribute their own time to these initiatives, as do the institutions that support the various scientists' work that eventually makes its way into the catalogue taxonomic backbone. All of this activity adds up to several million more Euros (World Register of Marine Species, 2017c), much more than any individual infrastructure's published core funding amounts. Adding a layer of complication to this schematic is the taxonomic work that happens *underneath* the Catalogue of Life within the GSD and RSD environments—all of which seeks

out funding for their own editorial and scientific work. What transpires from this narrative is a nested series of independently funded entities, beginning at the highest level in the Catalogue, cascading downward throughout the entire taxonomic system.

So when Doug Yanega points to *funding* as a key system underpinning this biodiversity and taxonomic work, he points to a rather convoluted network of funding agencies that are tightly integrated with the practices that structure science as a whole. In addition to this issue of funding is Doug's second major critique of the Catalogue: how are we to *review* the contents of the Catalogue to ensure data quality and completeness? How can we know what data is missing or correct, and how can we assess how complete any particular GSD database is within the system? One of the Catalogue's primary aims is to make a *complete* species list taxonomy that facilitates the global distribution and collocation of knowledge that helps us understand the extent distribution of biodiversity knowledge. The downside of this approach, of course, is that such a veneer of cohesiveness makes it difficult to *deconstruct* the component parts to reveal its inconsistencies and gaps.

Assessing data quality and completeness.

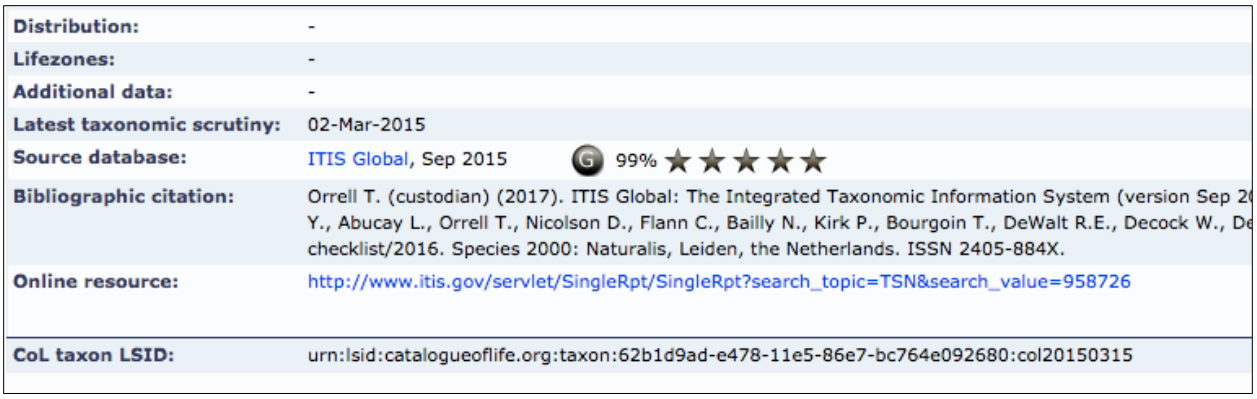
In order to assess *quality* and *completeness* in a taxonomic space, we must be able to assess data provenance and have general transparency about that data's known issues and shortcomings. The *true* extent of the effectiveness of systems like the Catalogue of Life is still unknown, and more importantly, we do not have articulated mechanisms by which we can measure this effectiveness. Certainly, one way we can assess quality is through the credibility and reputation of the institutions and individuals that provide information to the Catalogue.[98] Such credibility is made visible within systems by linking back to the contributing GSD or RSD,

[98] My previous discussion in chapter four regarding how Catalogue editors assess taxonomic contributions through illustrates this fact.

as well as the requiring the "Latest Taxonomic Scrutiny" data field for all contributed datasets to the Catalogue (Species 2000, 2014). This field group must include, the "name(s) of the taxonomic expert or editor, who is responsible for the taxonomic concept accepted in the source database and (b) date when the expert or editor [or small team] assessed the record" (2014, p. 10). Such linkages and attribution is essential in deconstructing the veracity of the Catalogue's individual entries, especially since the taxonomic editors depend entirely on the expertise—and reputation—of those that contribute this data. In addition to these provenance markers, source databases in the Catalogue provide a *confidence rating*, or dataset qualifiers (Species 2000, 2017c), for the taxonomic data, which is valued on the following scale,

1. Caution! This data set does not contain well scrutinised (*sic)* taxonomic checklist, and in parts may be a list of taxonomically unvetted names only. However, it is used temporarily by the Catalogue of Life to fill major gaps as only available source at the time. See database abstract for more details.
2. Caution! This data set is a scrutinised taxonomic checklist, but it is incomplete and at an early stage of its development. See database abstract for more details.
3. This is a well-scrutinised taxonomic checklist, but it is restricted to a subset of species by geography (regional database), or sector of biological discipline (e.g. thematic database in particular ecological area, conservation, quarantine, pest and disease control, medicine or molecular biology, etc). This data set was included in the Catalogue of Life to fill gaps at lower levels of the taxonomic classification (e.g. species, genera) as temporarily solution. See database abstract for more details.
4. This is a nearly complete and fully scrutinised taxonomic checklist with a good quality of expertise at the current stage of its development. 5 - This is a complete and fully scrutinised taxonomic checklist for an entire taxon with a high quality of expertise and frequent updates, which covers nearly all known species diversity in the taxon worldwide (Species 2000, 2014).

This information is then displayed in the main record at the appropriate taxon level (See Figure 23). Of course, the rating itself is provided by the contributing database, so the quality rating is dependent on the contributor's willingness and ability to acknowledge their own database's strength and limitations.



Distribution:	-
Lifezones:	-
Additional data:	-
Latest taxonomic scrutiny:	02-Mar-2015
Source database:	ITIS Global, Sep 2015 99% ★★★★★
Bibliographic citation:	Orrell T. (custodian) (2017). ITIS Global: The Integrated Taxonomic Information System (version Sep 2(Y., Abucay L., Orrell T., Nicolson D., Flann C., Bailly N., Kirk P., Bourgoin T., DeWalt R.E., Decock W., De checklist/2016. Species 2000: Naturalis, Leiden, the Netherlands. ISSN 2405-884X.
Online resource:	http://www.itis.gov/servlet/SingleRpt/SingleRpt?search_topic=TSN&search_value=958726
CoL taxon LSID:	urn:lsid:catalogueoflife.org:taxon:62b1d9ad-e478-11e5-86e7-bc764e092680:col20150315

Figure 23. Catalogue of Life taxon record for *Acidimicrobium ferrooxidans* Clark and Norris, 1996. Note the "Source database" field, indicating the ITIS Global source database, the database version date (Sept 2015), the percentage of completeness of the species list this entry is embedded within, and finally, the confidence rating for the quality of the taxonomic checklist (level 5 value).

Still other data aggregators, such as the Global Biodiversity Information Facility (GBIF), have attempted to articulate more quantitative and decisive mechanisms for quality assessment. Acknowledging the "potential problems related to incomplete or difficult access to all fields in its scheme," as well as in other "aggregated data research infrastructures," a task force focused on "GBIF Data Fitness for Use in Distribution Modelling (*sic*)" (Anderson et al., 2016) was established to articulate a series of recommendations on data fitness and efficacy for large-scale modeling use. Error rates are certainly an issue of great concern for the biodiversity community, illustrated by Robert Mesibov's (2013) audit of GBIF and Atlas of Living Australia (ALA) *Diplopoda*[99] occurrence records. His study indicated a high variance in data quality and completeness, and he ultimately questions the value of aggregated data repositories such as these for local purposes (2013, p. 16). One of the main recommendations by the GBIF Task Force centered on a users inability to accurately differentiate between quality and non-quality data:

> GBIF.org should serve indicators of precision, quality, and uncertainty of data that can be calculated practically, and preferably "on the fly", as well as summaries and metrics of completeness of inventories, at scales and for regions defined by the user. The summaries should display maps and graphs of completeness by region, time-period and taxa (2016, p. 2).

[99] The arthropod class, *Diplopoda*, consists of the millipede taxon group ("Current Status of the Myriapod Class Diplopoda (Millipedes)," 2007).

The task force specifically recommends clear fields to indicate error and uncertainty rates, as well as provide methods for users to visualize datasets to understand the larger contours of that data and "highlight possible inconsistencies and error" (2016, p. 4). GBIF is also transparent and forthcoming with the fact that their "interpretations" of the data can have unexpected errors.[100] "These [interpretations] do basic string cleanups but for many important properties we also use strong data typing. For example latitude and longitude values are represented by java doubles and the country, basis of record and many other terms which are based on a controlled vocabulary, are represented by fixed enumerations in our java API" (GBIF.org, 2016). To make any possible errors as a result of this process visible to users, potential interpretation error notes are appended to every occurrence record in the GBIF repository (2017).

Aside from data quality, Yanega notes that there is also the issue of being able to decipher the *completeness* of any given taxonomic hierarchy. As Alan Paton of Royal Botanic Gardens, Kew, remarked, we are now experiencing,

> That transition between how the data is curated at local level to truly global curation of data. And at the moment the emphasis is on local creation because we know it works and nothing gets missed …the next step is a community resource where people can begin to curate it and disambiguate species. But in order to get to that level you need a critical mass of data in one place. Because you can develop *all* the tools, but if there's no digital data correctly identified with identifiers with some level of authority (this is good, this is reasonable or not good), then...what would people do with that data? We are kind of in that transition and I don't know how long it's going to take us to [make] that transition. One of the main barriers to that transition is "do we yet have the critical mass of data digitized? We've got 10% of our collections databased [at Kew], 20% across the board. It means 80% isn't digitized. And that's just at Kew. Literature form 1924-2011, roughly, before born digital kicked in, is largely under copyright, and maybe 30% digitized (2016).

This snapshot of Kew's current state of digitization should be somewhat sobering when we think about the extent of biodiversity information available in any one of the iLife platforms we have been discussing. The problem is compounded by the fact that institutions all over the world experience this analog/digital divide. Digitization of museum specimen and type collections is a

[100] Certainly, the Catalogue of Life is also quite transparent about it's own limitations as well, which even a cursory look at their website would reveal.

costly endeavor—so much so, that most institutions cannot invest in these efforts. Some institutions, however, such as Naturalis Biodiversity Center in Leiden, Netherlands, have been able to invest a great deal of funding into digitizing their *entire* specimen collection. Such endeavors are an essential part of the taxonomic and biodiversity landscape, for the "transition" Paton refers to, depends on them. Naturalis recently completed a European-funded initiative to digitize their entire collection of specimens. The initiative itself was spurred on by a grant totaling €13 million to digitize *at least* seven million objects at the object level. To accomplish this feat, however, a tiered approach to the task was necessary: nine million objects of their forty-two million-object collection were selected to be digitized at the object level, while the remaining thirty-one million specimens were digitized at the drawer/box/bottle level. Boxes (or any standalone container) typically hold closely related species at a particular taxon level—say, genus (see Figure 24). As Renee Dekker, then Head of Collections, conveyed, these efforts have proven resourceful and effective in locating specimens rather quickly,

> Naturalis is a state collection, and we are the tenants of the state collection. They visit us every 2 years and they randomly select 44 objects and we have to find them within 10 minutes. Which is easy if you have only a thousand paintings; if you have 42 million objects it's a different story. But normally the scientists and the collection manager can link to that collection and find it within 10 minutes. The good thing now is that everything is databased at carrier level. I can find it and even the Director can find it because everything is [identified exactly] where it is in the collection—which floor, which cabinet, which aisle, which drawer, and in which position. And when the inspectors came recently, they randomly selected two microscopic slides —the smallest things, with many items in a box, and we have about one-hundred thousand microscopic slides—and because they are all linked and in the computer ... even I could find it, and not within ten-minutes, but five-minutes (2016).

A rather impressive feat given the size of Naturalis's collection.

There are two broad questions, then, to be posed in relation to *completeness* as it exists in the biodiversity taxonomic and data space: (1) how much data is digitally extant and available in the iLife consortium *at all*, and (2) how much of that digitally available data has been collocated appropriately in composite structures such as the Catalogue of Life. The former question is being addressed by Naturalis's attempts to digitize and move the intellectual objects of biodiversity

research online for broader global and systemic uses. While the latter question is our particular concern for this manuscript, the former issue cannot be ignored in the grander scheme of the discussion. The assessment and completeness of the data that populates databases like the Catalogue is one major concern, but so too is the very real issue of what is excluded from these systems for various reasons. A key question is how easily the digital surrogates created out of initiatives like those at Naturalis are finding their way into the systems, and perhaps more importantly, what information from these collections is not yet in a form amenable to global sharing.

Figure 24. Photograph of unidentified wasp specimen drawers at the University of California, Riverside, Entomology Museum. The orientation, structure, and detail of this specimen drawer are quite typical of those found at other natural history museums. Note the minute, multi-layered metadata labels that contain specimen-specific information such as unique institutional identification number, species information, etc. Photo by author, August 3, 2016.

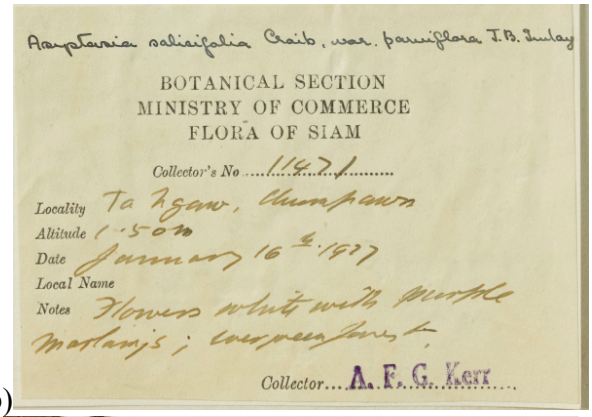
The fact remains that, even in the most highly funded and structured initiatives, full data integrity cannot be assured through any mechanism. Even Naturalis's extensive efforts did not include fully OCR'd label text for online searching capabilities. In addition, in this and previous digitization efforts, the notes attached to the labels themselves were not digitized, even though these labels hold vital information for scientists. *Every* piece of extant information cannot be transferred into the digital environment, and this will always be the case. A good deal of the

penumbra of metadata surrounding what we primarily understand to do the most important intellectual work in the production of species concepts—the specimen itself—is often excluded from the global systems. For example, oftentimes, much of the metadata from type specimen cards never makes its way into the global database environment (See Figure 25).

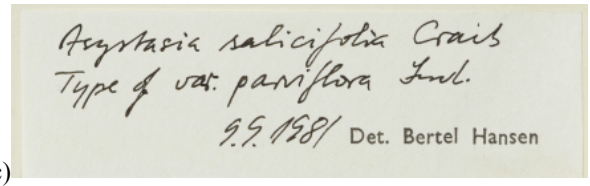
Determination (DET) labels are slips of paper that taxonomists paste onto the specimen cards over time to indicate changes that have been determined to affect the nomenclatural status of the specimen.[101] These slips serve as kinds of archeological markers for the historical evolution of the species concept. In the example in Figure 25, you can see that this specimen began as Asystasia salicifolia Craib var. parviflora Imlay, J.B., in 1981; was then changed to Asystasia nemorum Nees in 1982; subsequently determined to be Asystasia saliforlia Craib in 1983; and then reverted back to Asystasia nemorum Nees in 1998 (F). Kew's Herbarium Catalogue (2017b), being a particularly robust system, meticulously lists these species concept and nomenclature DETs over time (Figure 25, (g)), but many online systems fail to depict this level of detail in their online specimen card systems, losing vital information necessary to fulfill the goal of a fully-functional knowledge base necessary to answer large evolutionary or phylogenetic questions (see Figure 21 at the opening of this chapter). Such type specimen information is essential as well if a fully robust knowledge base system is going to serve as the foundation for quality evolutionary informatics work.

[101] I am thankful to Timothy Utteridge, Head of Identification and Naming and Senior Research Leader (Asia), at the Royal Botanic Gardens, Kew, for instructing me on the finer details regarding species concept and nomenclatural emendation practices.

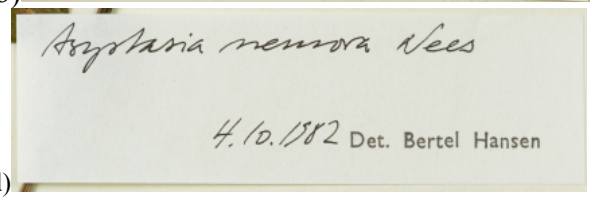
(a)



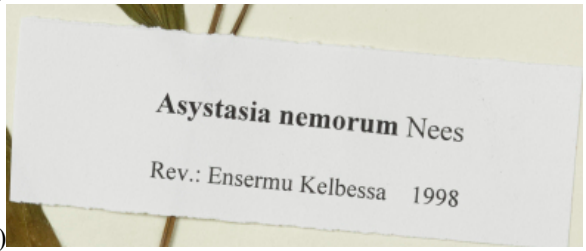
(b)



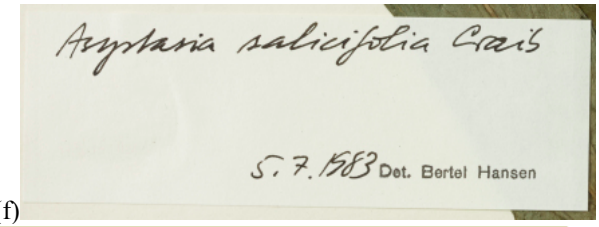
(c)



(d)



(e)



(f)

(g)

Determination History:				
Scientific Name	**Determiner**	**Determination Date**	**Type of?**	**Determination Notes**
Asystasia nemorum Nees	Hansen, B.	04/10/1982		
Asystasia salicifolia Craib var. parviflora Imlay, J.B.	Hansen, B.	09/09/1981	✔	
Asystasia salicifolia Craib	Hansen, B.	05/07/1983		

Figure 25. Specimen card and determination slip detail for *Asystasia nemorum Nees* (Royal Botanic Gardens, Kew, 2017a). (a) Full specimen card; (b) original determination label; (c) determination by Bertel Hansen, 1981; (d) determination by Bertel Hansen in 1982; (e) determination in 1983; (f) the most recent determination/revision by Ensermu Kelbessa in 1998; and (g) the database fields representing these changes.

But even as we gain the means, staff, and computational resources to digitize these various sources of biodiversity data, error is still an unavoidable aspect of this procedure. The *process itself* will inevitably introduce its own set of issues, as Yanega indicated to me, since computational systems are not yet as reliable as human readability: "Every conceivable mistake at every conceivable level will occur when relying on automated [digitization and georeferencing] systems,"

> It's hard to express the complexity of the problem in a single paragraph. There are two things: there's introduced error and there's original error. Original error is something a person is going to notice and figure out and resolve, and people tend to assume that original error is a very minor thing, and it's not. I've worked in more than a dozen different collections (not just our own, including the Smithsonian) ... [and for] all of them, roughly one out of every five labels has either an error or an omission that is significant in terms of when and where and who collected it. One of those three parameters was, in some respect, from something that is either misspelled, switched around, or absent (2016).

Introduced errors are errors such as incorrect coordinates (in the case of georeferencing systems) or OCR errors (in the case of text-based digitization). Validating and manually appending information to these digital sources becomes another essential phase of this process. Fixing these kinds of errors, however, takes time to accomplish, and until we reach the practically impossible point where all information is transferred into the database ecology, completeness will remain an elusive concept. Yanega's concerns certainly apply in specimen card spaces as well, for the transcription and/or OCR transfer of these texts into database fields is often fraught with error as well.[102]

While I have only presented a few examples, it should be clear that completeness of data in the biodiversity/iLife environment is judged relative to the availability of digitally-transferred data sets from analogue sources, as well as by the extent to which we can judge these sources to be vetted and curated to the best standards available. Not everyone is convinced that the

[102] One type of error encountered frequently is the inability to read handwriting on type specimen cards and/or specimen labels. I can personally attest to these legibility issues. While individuals intimately knowledgeable with certain collections may be able to decipher certain handwriting styles, hiring others to do so is often another way in which error infiltrates the database text space.

Catalogue of Life can gain a reasonable level of dependability. As was discussed in chapter four, the management taxonomy implemented by the Catalogue (functionally and intellectual separate from the process of taxonomic opinion) has been a core point of contention for the many individuals who do not agree with such pragmatic approaches to taxonomic management. The management hierarchy, and the mélange of taxonomies that constitute its structure, make its content somewhat tenuous and imbalanced, keeping certain communities skeptical of its value for local, taxonomically relevant reference purposes. David Patterson expands upon this notion,

> No there is still no consensus over how to handle the protists. Molecular analyses have tended to add a fair bit of noise to the picture, this has led to many speculations expressed in the form of classificatory structures, and the consequence is a lot of confusion. Some parts of the scheme appear to be increasingly robust, although the scope and definition of the taxa remain uncertain. Survivors at the top level seem to be the Opisthokonts (animals, fungi and close protistan relatives), Amoebozoa and Rhizaria. Excavates go in and out of favor, while chromalveolates and Archaeplastida are not solid. Similarly, at more distal points in the conceptual tree, some taxa, such as Chromists, are unsupported by much beyond wishful thinking and so are contentious.
>
> Ideally, the application of phylogenetic principles as criteria for retention or dismissal of taxa would be wonderful, and protistologists have been somewhat slow to move in this direction.
>
> A protist classification that is more consistent with currently available data can be found at eutree.lifedesks.org. It is a working structure rather than a reference structure. In that system, if relationships are unclear, the contestants for most proximate neighbors are placed as sister groups to minimize the risk of producing polyphyletic taxa (2009).

As was conveyed to me, it is much easier to assess the quality of an entity like a GSD, since these spaces are often tightly curated with—again—internally consistent methods and approaches. Such is the case with the protest databases references by Patterson, and any number of other group specific taxonomies. As Paton mentioned above, the transition to an environment where a "critical mass" of information is available is not yet upon us. A step toward this threshold is to ensure that completeness and quality of a system can be understood and transparent, so that they can be useful and reliable sources that scientific professionals can feel comfortable implementing as part of their daily practice.

But even exempting these issues, the Catalogue's taxonomy *is*, in fact, used by a number of data aggregators as a backbone structure. In what sense do these uses expose some of the Catalogue's *extensive* limitations?

The Catalogue's Extensive Limitations (or the Limits of Curated Spaces)

Issues regarding quality and completeness aside, it should be noted (and readily apparent from previous chapters) that the Catalogue is still a highly curated environment. Editors do their best to balance the needs of multiple (competing) communities to provide a functional snapshot of biodiversity life for more practical and pragmatic purpose. But even at over 1.6-million species, there are still many environments where this number is deficient for descriptive and organizational purposes. As we will see in great deal below, GBIF is one such example. One of the key extensive capacities of the Catalogue is its ability to serve as the taxonomic backbone for numerous systems across the iLife ecology and beyond. In practice, the implementation of its hierarchy has its limitations. In some cases, such as the Encyclopedia of the Life, the Catalogue is displayed among an array of taxonomies to show the divergent opinion of the taxonomic community (Encyclopedia of Life, 2017d). In others cases, such, as GBIF, the Catalogue is used as the core structuring agent for searching, browsing, and organizing its species occurrence records. The GBIF Nub taxonomy, as it is called, "is updated regularly through an automated process in which the Catalogue of Life acts as a starting point also providing the complete higher classification above families" (Global Biodiversity and Information Facility, 2016a). The case of GBIFs nub taxonomy is a fascinating one, particularly because it integrates the Catalogue of Life as its core, but in order to meet its site-specific needs, builds upon this core to cater to its diverse occurrence record data set. Expansion on this example is merited for it illustrates the potentiality and limitations of the extensive uses of the Catalogue rather well.

Before proceeding, it makes sense here to reiterate how GBIF functions differently in the iLife environment than the Catalogue. For one, the Catalogue's primarily currency is in valid and accepted *names* (as a nomenclator), and by connecting those tokens into valid taxon groups, produces a management hierarchy that can be used within other infrastructure to organize their data on top of this basic architecture. GBIF's main function, on the other hand, is to collect globally-produced occurrence data and provide "a single point of access (through this portal and its web services) to hundreds of millions of records, shared freely by hundreds of institutions worldwide, making it the biggest biodiversity database on the Internet" (Global Biodiversity and Information Facility, 2013b). Names and taxonomic relationships are the primary focus of analysis for the Catalogue, while in the GBIF environment it is the occurrence data that are the primary issues of concern, ultimately appended to a taxon within that hierarchy for searchability and organization. In terms of nomenclature, the Catalogue of Life provides a significant percentage of the nomenclature information for GBIF's Nub taxonomy, comprising some 3,175,925 names, or approximately fifty-four percent of the total GBIF namespace ("GBIF backbone taxonomy - Constituents," 2017). Even with this partitioning of responsibilities in place, GBIF is unable to take the Catalogue's management hierarchy and implement it wholesale. As Tim Robertson, Head of Informatics at GBIF, summarized,

> There is no global taxonomy that can organize the 730 million occurrence records. It just doesn't exist and therefore GBIF have had to assemble one. In the past, GBIF used to assemble this taxonomy from the occurrence records themselves. As records came in with kingdom, phylum, class, order […] and scientific names we would remove suspicious names, and try to assemble what was left. This proved to be too inconsistent and was messy. We found it needed higher quality source data. In 2011 we stopped that approach and we started building the backbone taxonomy using only what we believed were trustworthy checklist datasets. Using the Catalogue of Life as a basis, we integrated names accessed from nomenclatures like International Plant Names Index (IPNI) and Index Fungorum and others until we had enough coverage to organize the occurrence data.(2016).

As there is still no single taxonomy existing that covers all known names, GBIF is forced to build its own GBIF backbone on top of the Catalogue of Life out of operational necessity

(GBIF.org, 2016). Within this activity, a conflict thus emerges: all the 640 million-plus circulating occurrence records that find their way into GBIF collectively comprise an *undifferentiated* landscape of species concepts, data types, and taxonomic hierarchy formats. In GBIF "not every dataset includes information at the same level of detail," going on the presumption that "sharing what is available through GBIF.org is valuable, because even partial information answers some important questions" (Global Biodiversity and Information Facility, 2013a). Unlike in the Catalogue, then, the data coming in to GBIF is not curated to the same extent. We can essentially think of this data pool as *all the possible occurrence records for all possible species*.[103] The Catalogue's taxonomic space, on the other hand—for all the reasons we have discussed up until this point—does not yet have the entirety of the possible two million extent global species yet to be document and included.[104] While much of occurrence data ingested into GBIF can be matched to a *position* in the Catalogue's core taxonomy, enough of it lies outside of the Catalogue's curated core (recalling the Catalogue of Life Plus schematic) to merit building on top of that basic backbone.

Certain professional scientists, including Roderic Page, Science Director for GBIF and Professor of Taxonomy at Glasgow University, have indicated that perhaps GBIF should "take more 'ownership' of data quality, but that's politically tricky" (Page [@rdmpage], 2016b). Which says nothing of the fact that curating taxonomic and occurrence data is an incredibly time-consuming task. To offset this limitation of the Catalogue's hierarchy for their own purposes,

[103] In a taxonomic and data quality sense, we can think of this undifferentiated occurrence record space as being (somewhat) analogous to Bourgoin's notion of "names as strings," Stage 1, in his Global Names-Catalogue of Life Parameters schematic (Figure 12). In this space data is relatively unrefined and not yet formatted in ways that facilitate integration into the GBIF system.

[104] See Footnote 11, in chapter one, regarding the disputation of the two million species figure utilized by the Catalogue of Life.

GBIF began implementing its own Nub taxonomy, mediated by computational and algorithmic methods,

> We have algorithms that we have tuned over time… But we start with one checklist and overlay another and build it up progressively. We've been doing this for many years and because our work is very visible with occurrence data, we get a lot of feedback on issues. Whenever we receive feedback we add tests to the algorithm codebase so that they catch real observed issues and we actively fix those. We've refined this algorithm … [it's] not perfect ... but we believe we are getting better and better at it. What we have built now could potentially be an option for building the Catalogue of Life itself, or at least providing a sandbox of information for review through the CoL processes. The Catalogue of Life, by definition, is a collection of global species databases, plus a few regional ones that get integrated together—by using the integration approaches developed over time we might be able to accommodate more sources quickly. These could be considered provisional placement of names for final peer review and confirmation. …. The goal would be to build the names layer across nomenclatures and then start layering up the hierarchies of taxon concepts in the Catalogue of Life. Ideally these would be linked to literature... GBIF can provide the linkages to all the type specimen information … to start linking with the museums—it seems like it can be a more inclusive, complete Catalogue of Life. A challenge is getting all the people involved to buy into that kind of concept. (Robertson, 2016).

The situation described by Robertson is reminiscent of the Catalogue of Life Plus's model, where, instead of nomenclature alone, provisional *taxonomic positions* are produced and published, while feedback mechanisms are put in place to more finely attune the structure to current taxonomic opinion. Additionally, GBIF's algorithmic method is in-line with Yury Roskov and Frank Bisby's initial vision for how proto-GSDs were initially intended to function to fill the gap spaces of the Catalogue of Life's taxonomic backbone. Of course, as we have previously discussed, criticism over algorithmic approaches to building taxonomies are not universally accepted in the taxonomic community. Many scientists feel that such methods are not as refined as taxonomies assembled through individual mediation, in addition to the more political concerns regarding the substitution of computational methods for hard-earned taxonomic expertise. Of note as well, is that GBIF's taxonomy, like the Catalogue's, is meant to only be *good enough* for organizational purposes, while the refinement is left to experts that fall within particular domains. The Nub Taxonomy is just as *contingent* a document as the Catalogue is in practice.

The algorithms used to compile the Nub taxonomy and name checklist are freely available on github (Global Biodiversity and Information Facility, 2017b, 2017a), along with the database schema that connects these various names into a taxonomic schema ("gbif/checklistbank schema," 2017). The extent to which these taxonomy-building algorithms can remain effective is dependent and contingent on the quality of individual data sources. "Some data [GBIF does] have is poor (e.g., @catalogueoflife has mangled butterfly names)" (Page [@rdmpage], 2016c), in addition to the fact that—similar to the Catalogue's ingest of RSDs—many of the data sources are "[aggregations] of sources that [may] themselves be aggregations," making it incredibly difficult to fix the data at the source (Page [@rdmpage], 2016a, 2016d). Minute editorial miscalculations inherent in the Catalogue of Life have to be debugged and restructured downstream to meet various infrastructure-specific needs.

In addition to these limitations to the breadth of the Catalogue's hierarchy is its inability to quickly absorb new names as they are produced in the scientific literature. Due to the Catalogue's annual publication schedule (for the stable, Annual version), it takes a good deal of time for *new* names to enter the fabric of their management hierarchy. Even if contributing GSD and RSDs were to quickly integrate these newly published or registered names into their own databases at the local level, there would be a considerable lag time between that change and it's ingest into the Catalogue, its subsequent publication (according to the Catalogue's annual cycle), and then it's proliferation into related and integrated biodiversity data systems.[105] Such is the

[105] The registration of names in relation to publication is an evolving issue in nomenclature circles. When publications were produced primarily in print form, defining what was the unit of publication was a relatively easy matter. As publications have become primarily electronic in nature, nano-publications (Maddison, Guralnick, Hill, Reysenbach, & McDade, 2012) are making it difficult to decipher what counts as a publication unit. As one taxonomist indicated, "previously [the name] had to be published, but now, *electronic* publications must be both published and registered" (personal communication). The International Commission for Zoological Nomenclature (1999), for example, is now pondering a registration system in lieu of the publication requirement, such that names must be registered (and thus, centrally located for informatics purposes), but where publication will not necessarily be a requirement (especially since peer-review is not a *mandated* quality of a publication in the Code). "The

balance that must be made for all curated and controlled spaces. This poses a problem for outfits like GBIF who see the publication of a name as a moment where they can *both* add a name to its Nub taxonomy, as well as ingest the associated data for those publications into its database. Waiting for the appropriate taxon to show up on the Catalogue's radar is not the optimal approach for this purpose. These limitations have prompted some to call for "a more democratized [and] open Catalogue of Life" infrastructure, in which anybody can review the addition of new names, and where these newly produced names, often produced in micro-publications, can be ingested immediately.

Thus, GBIF is currently exploring ways in which immediate nomenclatural and taxonomic information uptake can be handled directly and automatically. Tim Robertson conveyed a story of how such a process might work in such an open system:

> There are also the new names being published though journals like Pensoft and through marking up journals through Plazi.org in Switzerland. They are some of the quickest people to publish their data sets of new names. Last year there was a new species of spider described and published in Pensoft—and by the time it was published, within about four hours we had it in the GBIF backbone taxonomy, with the type specimen [information]. It was quite exciting to see such a short turnaround – we had to do a lot of behind the scenes work to make that happen but it was nice to demonstrate that it was possible and what we are heading towards – "published to GBIF within minutes". Plazi and Pensoft in particular are pioneering rapid micro-publications (2016).[106]

Of course, the idea of a more "open" Catalogue of Life compounds some of the issues scientists have regarding the Catalogue: its lack of uniformity and its top-down approach to classification in general, to name only two. On the other end of the spectrum, such an open system can also facilitate ready and minute editorial changes to produce a more up-to-date system. Such systems

---

registration required [will] supersede the role publications used to serve" in this possible new workflow. This registration system would thus establish the link between name and type. I have been unable to decipher if similar initiatives are being pushed forward in the International Code of Nomenclature for Algae, Fungi, and Plants.

[106] Plazi (2017), ZooKeys (2017), PhytoKeys (2017), and the Biodiversity Data Journal (2017), were mentioned repeatedly in my interviews as key sources for *structured* rapidly-disseminatable publications that can easily be harvested and implemented in the biodiversity informatics environments such as those in the iLife consortium. ZooKeys, PhytoKeys and the Biodiversity Data Journal are all a product of Pensoft Publishers, "an independent academic publishing company, well known worldwide for its innovations in the field of semantic publishing and for its cutting-edge publishing tools and workflows" (2017).

would need to somehow incorporate peer review into the workflow so that taxa can be placed provisionally and then a series of experts could review and confirm that placement.

Error proliferation.

In addition to the ontogenetic transformations that take place within and without the Catalogue's taxonomic space, another issue identified in this composite and integrated space is the proliferation of GSD-introduced errors that radiate throughout the various networked systems of the iLife consortium (and beyond). This critique applies not only to the Catalogue's infrastructure but to *all* highly-integrated systems in general: these systems proliferate errors incredibly quickly given their interconnected nature. A discussion on the Taxacom list serve, initiated by Stephen Thorpe, offers one example of how errors (and in this case invalid author references) are introduced by taxonomic and database systems, and subsequently proliferated in the biodiversity online ecology:

> But they do [qualitatively transform data].
>
> Look at the authorship of Scolytus scolytus -
>
> In GBIF, EOL and Catalogue of Life 2007 the authorship is incorrectly listed a Wood and Bright 1992:
>
> GBIF: http://data.gbif.org/species/14616352/
>
> EOL: http://www.eol.org/pages/691357
>
> Catalogue of Life 2007: http://www.catalogueoflife.org/annual-checklist/2007/show_species_details.php?record_id=4242138
>
> However, in ITIS the authorship is corrected listed as (Fabricius, 1775).
>
> The problem began when a mash up was made from the Electronic Catalogue of Curculionoidea website.
>
> It correctly listed the authorship of Scolytus scolytus and cited the publication of Wood and Bright 1992 as the source. Somehow, the mash up dropped the authorship name and replaced it with the citation name. Then it spread ....
>
> Now, almost every weevil that occurs in North America and was listed in the Wood and Bright 1992 publication has Wood and Bright as the author of the those species: Here is the EOL Scoltyus species list. Run your eye down the list to

> see how many species have Wood and Bright 1992 as their authorship: http://www.eol.org/pages/49702
>
> I sent emails to GBIF and EOL without receiving a reply and so like hitting your head against a brick wall — I felt better when I stopped.
>
> I am beginning to wonder whether discrete taxon treatment websites are indeed better than those that attempt to do all (Walker, 2010).

This post highlights a number of issues that many scientists feel are major drawbacks of distributed data systems: the inability to validate and edit small bits of big data systems; the speed with which these potential errors are distributed; and, finally, the lack of feedback mechanisms for systems such as the Catalogue of Life, GBIF, and attendant systems, to remedy errors that are identified by specific users and user groups.[107] On the one hand, this is a problem of data management at the stage of collocation: the necessary steps that need to be taken to establish the veracity of any particular source may have been overlooked by either the database provider or an editorial board. But as we have seen, these systems are working with complex data sets, and there will be no situation in which 100% accurate data will be reached through the work of any editorial body. Using the example above—and assuming that *Scolytus scolytus* Fabricius, 1775 is, indeed, the correct nomenclatural form—Catalogue editors had to make a good-faith decision at some point to include the "provisionally accepted name" from the "WTaxa: Electronic Catalogue of Weevil names" database (Species 2000, 2017e), and *not* from the ITIS database, which incidentally, still correctly lists the *Scolytus scolytus* Fabricius, 1775 form of the name (ITIS, 2017b).[108] It could be the case that the "Electronic Catalogue of Weevil names" was more comprehensive in those taxon groups over ITIS, and so what the Catalogue

[107] To be fair, the Catalogue and GBIF *do* attend to recommendations and suggestions submitted to them directly. It is often the case, however, that they are not contacted regarding these issues, or individuals do not know the most effective mechanisms to notify them of such issues.

[108] The Catalogue of Life and ITIS still have the "errors" described in Walker's Taxacom post as of March 1, 2017. It should also be noted that I take no position on the correct form of this species. I only use this example to show that discrepancies exist in this informatics landscape and that these issues proliferate quickly in this online environment.

gained in terms of species coverage and breadth may have outweighed any small errors that the database introduced into the system.[109] On the other hand, this is a theoretical issue whereby the ontological instability of biodiversity database systems comes into direct conflict with the practice of *using them*.

Much of these issues can be rectified with an effective feedback mechanism, and as has been illustrated, museums, taxonomists, and databases, do, in fact, *want* feedback. But even as individual systems add mechanisms to make requested changes within their databases, they are likely to get wiped out as these systems transfer data between them over time. And all of this assumes that there is, in fact, somebody to contact to fix the error. Recall that GSDs are the source of most of these data points. And also recall that these GSDs are often manufactured and updated by limited staff—and in some cases, only one individual. An increasingly difficult problem in these spaces is how to manage and curate data sets that have been orphaned or abandoned by their creators? As was conveyed to me,

> A lot of people using data find issues and we aim to mediate their feedback back to the data publishers. We're beginning to recognize—we have to recognize— that some datasets are actively curated and some data really will never be touched again. As an example … the museum community they really do seek feedback—they curate their specimens, and they curate their databases… If end users have feedback - even tiny little pieces of info – it is often acted upon. Then there is a whole class of content originating from citizen content initiatives; those are very active (iNaturalist, for example), and want and act on feedback. Then there is survey data, machine generated data, content where people have produced for project then moved on in their career – this is often content which will never be visited again by the originator. We need to consider how we treat edits that should be applied to this data. During 2017 GBIF aim to identify those cases and develop a community curation approach whereby people can correct issues in these datasets and republish a new version (Robertson, 2016).

Additionally, in algorithmically constructed taxonomies—such as GBIF's—errors like this may be a product of computational error. As Nico Franz stated, "But that means there are no primary authors [for the GBIF taxonomy]? No people to directly dis-/agree with" (2016). One problem here is finding the correct source to contact to rectify these errors is equivalent to hitting a

[109] I do not know the reasons for the inclusion of the "Electronic Catalogue of Weevil names" database over ITIS, this is an example purely for analytic purposes.

moving target. So there are two *kinds* of data that exist within GBIF—and, by extension, any system that aggregates disparate data sets—data that is active (in that there is a responsible body concerned with the continued curation of said data asset), as well as what Tim Robertson called, "immutable" datasets that have nobody in charge of them and able to mediate incoming feedback. The end result of this bifurcation is that datasets need to be identified according to a set of activity parameters, and if it is categorized as immutable, mechanisms need to be put in place so that somebody can take control of managing that data and curating it to meet contemporary science and discovery uses.

Divergent traditions and nameless taxa.

The final limitation of the Catalogue that will be discussed here is its inability to absorb and formulate *positions* for classificatory productions that *do not* conform to Linnaean-formulations. The increased use of genetic markers in the form of DNA barcodes (such as the mitochondrial C01 gene sequence) have been increasingly useful in constructing phylogenies (Waterton et al., 2013; Erickson & Driskell, 2012). The outputs of such genetically-based phylogenetic examinations are tree drawings that depict the hypothesized relationship between various entities (not unlike the output of numerical taxonomy discussed in chapter four). One result of the increasingly popular approach of phylogenetic inference is the "proliferation of taxonomic categories" (K. de Queiroz & Gauthier, 1992, p. 457). DNA tends to "split" more than it does "lump" species together into taxon groups. The applications of names to these DNA barcode strings, however, and the classification of these categories within existing taxonomic schema, are entirely separate activities performed *after* these phylogenies have been constructed. Increasingly, names are not being applied to this growing cache of genetically labeled

information. An interview with Berry van der Hoorn, Group leader for Biodiversity Discovery at Naturalis Biodiversity Center, provides an example of such a practice:

> So we're working, for example, on water quality, taking water samples from a ditch somewhere and then we extract DNA from the sample to identify species that are living in that environment. And this has all kinds of applications: for impact assessment, for agriculture, for deep-sea mining … wildlife forensics, food safety, etc. We have been running a project where we barcode a lot of species around here. I think we've barcoded 40,000 specimens from our collection and have a lot of [newly collected] material. This water project is able to function … because we have libraries of barcodes, so [this project] has a huge impact on taxonomy. It's an extra source of information for the real taxonomist, so if you want to do a revision on a certain species groups, if you are very old fashioned, you only look at morphological characters, but that is [less practiced] these days. You should … take molecular traits into account.
>
> Barcoding works quite well. For example, you don't even need species names anymore. Sometimes, for us, it depends on the questions you ask. We [went on] an expedition [to] a small island in the Caribbean … to check how [species] diversity varies over the island. Sometimes you don't even want to know the species name, you just say, we found 150 spiders here, 12 spiders there, and then that's enough. They call the output of this "OTUs"—unidentified, operational taxonomic units. And you can use [these OTUs] fairly well as a biodiversity index and then you don't even need to recognize the species itself. And of course, you have to take into account that you miss some species because you cannot identify some species using the CO1 gene, for example. But that doesn't matter for this biodiversity index because … for defining the water quality it doesn't really matter. It's the one species or the other one—these are minor details.
>
> Many taxonomists see this [approach] as a threat, but for me [I think] it supports taxonomy because it gives it so much new energy and information. You can build up so [many more] applications than you could before. We show how much taxonomy is [still] needed (2016).

Interestingly, as van der Hoorn conveys, the divergence between traditional taxonomy (and the application of names to taxon) and the growing interest in barcoding as an approach to constructing taxonomies, is a matter of asking a different *kind* of research question. For van der Hoorn, the question was ecologically based, formulated within and for very specific conditions. The application of names is secondary to solving a project- and funding-specific research query.

In practice, numerous scholars have pointed to this widening divide between traditional and phylogenetic approaches, and the detrimental effects it is having on the adequate accumulation and collocation of scientific knowledge as each proceeds forward invoking and implementing different methodologies. Nico Franz (2005), at the University of Arizona, documents this increasing tendency to not translate phylogenies into classifications in his article "On the Lack of Good Scientific Reasons for the Growing Phylogeny/Classification Gap." Franz

notes, "by supplementing a traditional classification with a more precise estimate of phylogeny, one has not yet 'removed the need to use' any or all parts of that classification. In the vast majority of cases, the more recent phylogenetic analyses are properly considered revisions of pre-existing hypotheses (however coarse) about the relationships among taxa and the evolutionary histories of character traits" (N. M. Franz, 2005, p. 496). The traditional modes of taxonomy that have been built over the last 250 years, including the application of names to taxon groups, are essential to contextualizing and making meaningful the results of phylogenetic analysis. Additionally, unnamed phylogenies become siloed from this cache of knowledge linked to the historiographical record. One camp cannot communicate with other, thereby limiting the ability for systematics as a whole to proceed forward as a coherent unit and to build upon the virtues of each approach.

Despite these competing views, there is certainly a synergy between the two camps that can flourish. A project such as the one described by van der Hoorn provides the raw data that taxonomists can use to produce more robust and complete classifications,

> When we barcode we intend to [eventually] upload it to [the Barcode of Life Data System][110] to stimulate all this taxonomic research. We don't have the capacity ourselves to name these organisms. That would take too much time. But if we barcode a lot of stuff and we give it back to [taxonomists] to sort it all out, work on it, describe the species, and give them names, that would be great. We give them a web platform to publish that information and contribute to these species catalogues, [and they could potentially] get rewarded professionally [by citing this work] (Van der Hoorn, 2016).

In the meantime, however, while these increasing caches of barcodes are being produced, this information fails to make its way into aggregated and composite systems like the Catalogue of Life. The Catalogue, as previously discussed, has created the Catalogue of Life Plus as an infrastructure to take these OTUs, and eventually give individuals feedback mechanisms by which these numerical name-place-holders can subsequently be described according to particular

[110] For more information on the Barcode of Life Initiative and Data System, see (The International Barcode of Life Project, 2010).

codes of nomenclature. Such Catalogue of Life Plus infrastructure can support the synergistic activity described by van der Hoorn. But this is a far cry from being able to take these non-nomenclatural values and add them into the management hierarchy for inclusion into its checklist and subsequent re-use in the iLife system.

GBIF is working on mechanisms by which genetic barcodes and other kinds of similar data can be appended to their Nub Taxonomy framework. For example, species that are databased can often be recognized at a higher-taxonomic level—say, at the order or family level. For example, in Berry van der Hoorn's described project above, spiders were collected from an area through the use of passive traps. GBIF's goal in this case would be to append the barcoded data *at the highest known taxon level*—in this case, the barcoded spider data would be appended to Araneae order level of the Nub Taxonomy (See Figure 26).

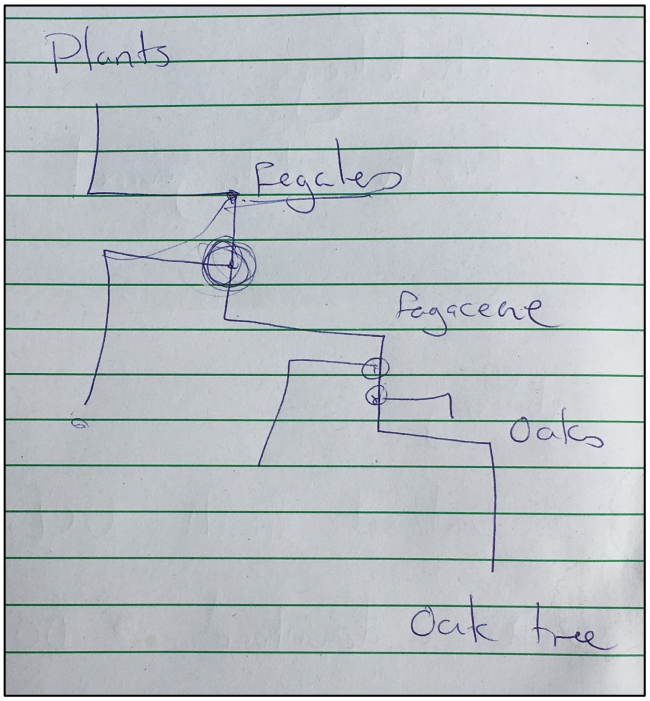


Figure 26. Hypothetical GBIF taxonomic hierarchy scribbled on note pad during an interview. The circled nodes on the taxonomy indicate the insertion of genetic or other non-Linnaean-named data at the lowest taxon level possible to suitably handle the data (personal communication, September 19, 2016).

The assumption here is that the availability of data imperfectly placed is preferred over perfectly situated data, given that the latter can take a great deal of time to assess. Users seeking out this information have the ability to search through the GBIF portal to locate this information. A fundamental information problem here, and one which Information Studies should be closely attuned to, is how the application of computational or statistical methods is only a small part of the overall goal of classification systems. Quentin Wheeler states, "phylogenetic classifications are optimal for storing and predicting information, but phylogeny divorced from taxonomy is ephemeral and erodes the accuracy and information content of the language of biology" (Wheeler, 2004). Language—and a set of name tokens, to be more specific—is essential as the concatenating force to bring together knowledge from different scientific sources. The biodiversity informatics world is closely attuned to these issues and divides, and are attempting to build classification systems that are more inclusive of multiple approaches to the production of classification systems. The goal of taxonomies such as the Catalogue of Life are to create inclusive systems that can, potentially, act as platforms by which the knowledge contained within multiple classifications can be distributed across the globe for use in any number of domains.

Conclusion

Referring back to the epigraph by Patrick Wilson that opened up this chapter, there are a number of dimensions of a documentary instrument that can increase or decrease an individual's degree of exploitative power: the extent of documents within the bibliographic universe over which they have control; the range of activities that can be performed on those accessible documents; and, finally, the extent to which you can 'own' the object such that it can be manipulated or re-articulated for individuated functions. As a retrieval-oriented system, the

Catalogue has set out to maximize all of these valences: to record as many biodiversity concepts-as-documents as possible; to make these documents as combinatorial and useful as possible as a base for new knowledge; and finally, to allow the reuse of its taxonomic document for a multitude of *extensive* uses. In the process of such exploitation and extension, however, both the GSD taxonomies and the Catalogue itself, change in constitution, contingent as they are in practice. Joseph Tennis's second-order classification theory was invoked to frame this change. Finally, some critiques of the Catalogue's composite form were offered to illustrate its extensive limitations within the biodiversity iLife environment. What these critiques offer are ways in which we can conceptualize the structure of the Catalogue as pushing against some fundamental social, cultural, and structural scientific traditions that are deeply engrained in taxonomic and biodiversity culture. Operations that are somewhat clear and more easily controlled within traditional internal-coherent taxonomic structures—data quality, articulation of taxa completeness, errors reconciliation, etc.—are a far more difficult to handle and process in composite structures. Yet without such aggregating mechanisms, the full extent of our exploitative and extensive power over our *collective* documentation and knowledge is limited.

In Henry Bliss's, *The Organization of Knowledge and the System of the Sciences* (1929), he exalts the "order" of scientific classification, arguing that the "practical classifications" of libraries,

> Do not conform to the scientific order. They have been constructed by those who did not rightly apprehend that order or who ignored it. The foundations of those systems were laid a half-century ago, or nearly, when the order was less clearly established in the consensus. Their makers were intent on constructing *practical* classification that did not see that the better classification conforms to the system of science the more serviceable it will be" (1929, pp. 411–412).

The turn toward composite and management taxonomies in scientific practice represented by the Catalogue—quite 'practical' and pragmatic in nature—troubles the rigid consensus-based order Bliss expected of the systems we create. As we have seen, however, biodiversity science is

anything but at a state of consensus; it defies universal consensus, for such consensus would indicate and scientific stasis. The Catalogue's pragmatic approach does not purport to mimic any scientific system, but rather chooses management and access as the primary function over fidelity to any given scientific argument. Opinions as to the overall effectiveness of the Catalogue may differ, but one fact is sure to remain: composite systems are increasingly necessary for switching mechanisms in a world filled with fragmented data sources. While description-oriented taxonomies will always remain a fundamental instrument in biodiversity work, retrieval-oriented mechanisms will continue to be a vital tool, connecting that data to the larger, diverse, global pool of scientific discourse. Identifying the downfalls of systems like the Catalogue is essential if bases of knowledge are to provide accurate foundations for our most pressing global biodiversity questions. As Patrick Wilson indicates, the success of any documentary instrument as a means of access is entirely dependent on our ability to understand its limitations.

As systems like the Catalogue forge forward in their attempt to concatenate global data, they will continue to balance the needs of local, specialist-specific scientific practices with generalized data use. This is a core tension and conflict within biodiversity informatics spaces, as Bob Mesibov explains,[111]

> The core issue here is the spectacular disconnect between working taxonomists and the acronymists who want to be the primary interface for access to specialist-produced data. Most specialists aren't asking 'larger questions' because they're thoroughly occupied with smaller ones. They don't need the infrastructures, but the infrastructures not only have an absolute requirement for specialist-produced data, but wistfully hope that specialists will assist with data quality within the infrastructure. Note that the acronymists aren't handing out money to assist the specialists, either with data quality work or with data generation. These are almost entirely separate enterprises. The argument 'We're all working on the same project and should support each other' doesn't wash with me, and won't wash until I see the pattern of growth in acronyms and decline in taxonomy start reversing (B. Mesibov, 2010).

[111] As of 2010 Bob Mesibov was an Honorary Research Associate at the Queen Victoria Museum and Art Gallery, School of Zoology, University of Tasmania.

Abbreviation, however, does not necessarily have to be equated with lesser quality. Systems like the Catalogue are continuing to develop and enhance their documentary properties. As Alan Paton gestured toward above, the science of taxonomy and biodiversity informatics is in a transitory state. There will be a point when a critical mass of data is obtained (even if we will never have it *all* at our digital disposal, for new species are constantly being discovered and the species concepts that represent them are contingent in nature). As CERN-like arrangements are solidified in the biodiversity community, infrastructures like the Catalogue will only improve as more-and-more knowledge is ingested into its database. And as evolutionary informatics continues to harness the potential of these aggregating infrastructures, the value of the Catalogue will become ever more apparent and necessary in structuring that collective knowledge. Full consensus on the Catalogue's function may never be present, but over time, perhaps this "disconnect" described by Mesibov can be bridged with fully integrated online systems that can meet the needs of many communities. It is really only a matter of time, mutual professional trust, and collective vision.

Chapter 6: Conclusion: Contingency and Future Trajectories

When the Global Taxonomic Initiative (GTI) was proposed as part of the Convention on Biological Diversity (CBD) in 1998, there was a keen awareness of the importance of taxonomy to the production of good biodiversity science broadly conceived. Without taxonomy there is no way to organize the species and taxon specific knowledge in any useable and shareable way. As stated by the CBD, "The lack of knowledge of key groups of organisms of importance to humankind, many of which have global or multiregional distributions, calls for a global dimension to taxonomic activities" (Convention on Biological Diversity, 2003, p. 24). The Catalogue of Life arose, in part, to help meet this need to understand the extent and depth of *all* taxonomic data—past, present, as well as to provide a mechanism by which future knowledge forms can be integrated into this structure. Such an overarching and global view entails a great deal of coordination at many scales, that includes both a focus on minute nomenclatural forms, as well as grander visions about how taxonomic knowledge can be tracked over time. It is this coordination that has been the main subject of this entire manuscript.

More specifically, this dissertation set out to answer two very basic lines of inquiry: (1) how can biodiversity taxonomic practices inform our notions of information, documents, and concept representation within the discipline of Information Studies, and (2) how might we understand *composite taxonomies* as information systems designed to *both* represent biological knowledge *and* coordinate efficient data communication? In general, biodiversity taxonomies can inform the discipline of Information Studies in a number of ways. For one, they shed light on the documentary control of concepts and objects (species and species concepts) that are grounded in some in natural phenomena—phenomena we can test and measure such that these empirical observations help us better articulate their specified *position* in a full classification of

biological life. But despite this referential 'ground' of external objects, consensus on their final conceptualization are anything but straightforward. Species concepts are, by definition, contingent—the practice of biodiversity science, in fact, is defined by these multiple and idiosyncratic taxonomic arguments. *General* consensus on taxon's position can generally be held, but even the most obvious taxonomic placement is vulnerable to rearticulation.[112] New methods are employed, new theories are developed, and new data is always being collected. The description of *texts* and *works* in bibliographical systems—and their subsequent classification—however, is much more subjective in nature: cataloguers are always seeking out the best decretive mechanisms to facilitate the use and needs of a user; users change, however, as do the contexts in which we need one form of information over another. Such description is predicated on the way somebody sees the world, and any other person can see things quite differently (Furner, 2009, p. 9).

So, at the heart of this manuscript is a deceptively simple proposition: all knowledge organization systems are constructs of cultural and historical circumstances, manufactured as artifacts of certain spatiotemporal positions. This, in and of itself, is not groundbreaking—Hope Olson (2002) articulated all-too-well the subjective nature of description and its subsequent properties of power in social conditions. What the Catalogue of Life exemplifies is a taxonomic space that is attempting to embrace diverse taxonomic conditions within the boundaries of one, multi-valenced taxonomic document.[113] The theoretical implications of this move for Information Studies are fundamentally significant: within a space of strong taxonomic opinion,

[112] On March 23, 2017, a new hypothesis for the classification of dinosaurs was released by University of Cambridge scientists (Baron, Norman, & Barrett, 2017). The longstanding classification between Ornithischia and Saurischia, held by consensus since 1888 (Ghosh, 2017; Wade, 2017), has been upturned based on new phylogenetic research of 74 taxa and 457 traits (Baron et al., 2017, p. 501).

[113] To be sure, I am in no way intending to convey that the Catalogue is perfect, no system ever can be.

where taxonomies have long been held as the primary mode of arguing serious scientific questions, the Catalogue has dedicated itself to changing the functions and epistemological foundations of *what taxonomies are supposed to do*. For the Catalogue, the exemplification of some 'real' natural world circumstance is less important than representing the *diversity of opinion* on that matter. What the Catalogue loses in internal consistency, it gains in cultural and historical breadth. What it loses in taxonomic specificity, it gains in historical perspective. The rhetorical move in embracing such an approach seems to me particularly powerful. And of course, the existence of the Catalogue does not negate the *absolute* necessity of descriptive-oriented taxonomic systems to biodiversity practice—these are fundamental scientific instruments and technologies. It merely annunciates the equally necessary position that *switching* between these divergent opinions in an equally accessible taxonomic space is also important. There are other questions and concerns that must be attended to outside the confines of our individual research questions.

Imagine if such switching mechanisms existed to bridge epistemologically and ontologically distinct cultural KO systems of Western science and indigenous Native knowledge sets (or any other distinct system, within any discipline, for that matter)? Would that switching system be perfect? Certainly not. I am not even sure what would constitute perfection in this space. What such a system—with all of its flaws and successes—would declare, however, is that such conversation is important—that the aim of diverse and inclusive approaches to knowledge organization is a fundamental tenet, not only of our theoretical discourse, but of the *practical* mechanisms and instruments we design that serve the direct needs of these various constituents. Such a system would state that the drive for a perfect and consistent system is antithetical to the ways in which information, documents, and knowledge function as social, historical, and cultural

constructs. It means that well-meaning and imperfect mediation is more respectable than none at all. If we take classification to be a process, then we also understand that classification can take many forms and occupy many different kinds of spaces.

Zooming out a bit, this manuscript also attempted to bridge the concepts of information and documentation in Information Studies with the practices and theories of biodiversity studies. I wanted to take foundational texts by Patrick Wilson and A. Broadfield, among others, and apply them to new and little examined spaces concerned with the documentation practices of the biodiversity and museum sciences. How do concepts of biodiversity evidence function within the representational spaces of the taxonomic document? And how can such an approach add to our current understanding of how document concepts are constructed within the field of I/S? At this point, it might be useful to rearticulate the general trajectory of this manuscript, if only to more clearly articulate how these questions, and many others, have been directly addressed.

Documentation and Document Contingency

This dissertation's first task was to more broadly define the concept of knowledge organization to be more inclusive of extra-disciplinary approaches. To more broadly define the scope of our notion of documents and concepts, to be more inclusive of those practices we see in biodiversity studies. I argue such inter-disciplinary approaches are not only essential to sustaining momentum within the information disciplines, but that disciplines not otherwise acquainted with Information Studies literature can gain from our theories and historical knowledge. I also sought to bring a few sub-disciplines in Information Studies into the conversation with one another that, I believe, are central to the space of classification and knowledge organization: bibliographic studies, documentation studies, and philosophy of information. Thus, in chapter two, I established the documentary universe of biodiversity studies,

illustrating the primary kinds of documents-as-evidence that constitute the taxonomic landscape, including information, data, and documents, and how these fundamental concepts work within biodiversity systems to produce progressively more functional and truthful units of knowledge for use in scientific practices. *Database-documents* are also established as contingent documentary spaces that consist of compiled evidentiary forms (the represented documents) and emergent forms (texts that constitute the Catalogue's space as a fixed and dynamic documentary resource). These emergent forms broaden our understanding of Wilson's notions of *texts*, *exemplars*, and *items*.[114] More fundamentally, though, examining a database as a document itself helps us better articulate how documents are not static entities, but are rather *contingent*, combinatorial structures that situates one of the basic truths of knowledge organization: the objects and subjects that we strive to represent, describe, and locate within systems are un-fixable, and thus the role of the knowledge organizer is to embrace this fluidity and craft systems that can embody dynamism and fluid representations.

This landscape then sets the stage for my subsequent problematizing of how these document entities are represented in databases via nomenclatural text strings, and how *control* is defined and managed in these contingent spaces. Chapter three expands upon the notion of *unruly* and *complex concepts*, a way in which we can understand the shifting representational and evidentiary notion of the species concept. Species concepts are understood as the triangulation of three entities: nomenclature, type specimen(s), and biodiversity/taxonomic literature. The relationship between these three evidentiary structures (kinds of documentary *warrants*) is constantly shifting and redefined over time, rendering any documentary arrangement continually contingent as new research and evidence is brought into examination.

[114] Such a conceptualization of these entities also brings into conversation elements articulated in the schema, Functional Requirements for Bibliographic Records (FRBR).

These *unruly concepts* can only be represented by text strings in the database environment, so an extensive discussion is had regarding how nomenclatural control is attained, and the kinds of problems shifting nomenclature presents to the study and assessment of authoritative taxon groups. Crucially, this kind of *control* marks a definite shift from *descriptive* to *exploitative* power—as names are disambiguated and validated in the database space, they become progressively more knowledge-based, and thus become more functionally *useful* as part of the biodiversity environment and within scientific discourse. The Catalogue of Life Plus is then introduced as an organizational and workflow mechanism to control for the proliferation of name-token types in order to *fix* these complex concepts in temporarily static forms. Such an examination, it seems to me, unearths two important notions about *concepts* and *subjects-as-concepts* in knowledge organization schema: (1) that their authoritativeness and usefulness as valid forms of discursive knowledge are culturally-defined and continually shifting; and (2) that the biodiversity environment has begun to think about mechanisms by which such epistemological and definitional change can be charted, documented, maintained, and negotiated, over time *without reifying one concept form/representation over any other*.

Taxonomic Contingency and Extensive Flexibility

The second line of inquiry in this manuscript is to assess consensus modes of classification as spaces that force two (potentially) competing modes of organization together: classification as a heuristic, descriptively based system, and classifications as data management, retrieval-based tools. The Catalogue is presented as a mediating platform where numerous epistemologically distinct classification systems intermingle for the purposes of data management and collocation. Traditional classification systems in the biodiversity sciences are typically understood to be internally consistent ontological constructs: clear operationalizations

are articulated for what constitutes the order of knowledge, classes, and relationships that are generally predetermined.[115] And such specifications allow a user to easily negotiate and understand the taxonomic instrument's logic and capability. Consensus structures such as the management classification of the Catalogue of Life, on the other hand, disrupt this notion of internal cohesion. Chapter four, therefore, broadens our previous discussions to include the *documentary instruments* themselves: the taxonomies-as-knowledge that organize names into meaningfully related taxa in systems. I examine the *specifications* of the documentary instrument, one of the most important being the ability to understand and deconstruct its arrangement and organization. I describe two kinds of biodiversity taxonomic instruments: description-oriented and retrieval-oriented. Examples of the former, descriptive-oriented instrumentation are taxonomic models that represent internally-coherent *arguments* about scientific knowledge—this is normally what we think when we conceive of biological taxonomies: phylogenetic models, for example, that tell us some consistent and unified story about evolutionary relationships. A management classification, however, is an example of a retrieval-oriented system, designed to access data that might be separated from its ontological and epistemological point of origin. A consensus structure is the culmination of taxonomic coordination that takes Global Species Databases (*inherited* instrumentation structures) and merges them together to gain a better understanding of *total* extent knowledge about global organisms. Such management structures can then be used to organize data on a global level. This capacity for re-use, I argue, expands Wilson's two existing *powers* of bibliographic control—descriptive and exploitative control— with a new kind of *extensive* flexibility. The designed

[115] Even in faceted systems such as S.R. Ranganathan's (2006) colon classification had pre-determined ontological categories established, such as facets and classes, that dictated the application of terms to described objects.

extensive nature of the Catalogue uniquely positions it to radiate its influence outward in the iLife consortium and beyond.

Finally, chapter five takes these contingent taxonomic document forms, examines their systemic purposes, conveys how they can be used to produce and trace historical taxonomic trends as *knowledge bases*, and examines their extensive limitations. Using the Catalogue as a kind of historical taxonomic concept repository can potentially help scientists make larger claims about macroscale evolutionary and phylogenetic questions.[116] Questions such as these leads me to discuss the ways in which taxonomies *change* over time, and how such changes can help us theorize and understand the *extensive* potentials and limitations of structures like the Catalogue. However, for all of their data-management virtues, the Catalogue has not yet been universally accepted in scientific circles as the most productive route for taxonomic science. Reasons why this might be the case are enumerated, and include: a distributed and unsustainable funding model (arising from the essential funding structures that support scientific work in general); given the mélange of databases that inhabit the Catalogue's space, it is difficult to assess its *completeness* and *quality*; the heavily-curated nature of the Catalogue makes it difficult to quickly embrace new taxonomic or nomenclatural knowledge; its interconnectedness in the iLife system proliferates errors that are not easily fixed; and finally, given that the Catalogue uses Linnaean species names as the collocating element, finding ways to include *nameless* taxa, such as DNA barcodes, is especially difficult. What we see here, are that the tensions between descriptive-oriented and retrieval-oriented approaches to classification systems have not yet been ameliorated. The reasons for such a divide arise from complex and fundamental social, cultural,

[116] Such an approach is often called "evolutionary informatics" (R. Page & Michener, 2012; Parr, Guralnick, Cellinese, & Page, 2012).

and structural scientific traditions that are deeply engrained in taxonomic and biodiversity culture.

Coming full circle, then, this manuscript examined how documentary *control* functions within an environment essentially defined by *contingent* concepts and documents, and how the disciplinary conditions within the biodiversity sciences are negotiating this tension through the basic knowledge and organizational structures that act as the fundamental instrument for taxonomic activities: classifications.

Future Trajectories: Alternative-Synthetic-Classificatory Examinations

Now that this first phase of my research has ostensibly come to an (artificial, but necessary) endpoint, I want to be able to articulate how it is that I conceive of going forward with this theoretical foundation in place. To explain, I want to return to the fourth stage of Thierry Bourgoin's "Global Names-Catalogue of Life Parameters" introduced in chapter three of this manuscript (Figure 12), focused on the entire nomenclatural and taxonomic context that the Catalogue of Life is functioning within. Recall that Bourgoin provides a broad overview of how undifferentiated taxon names get slowly transformed and mapped into taxonomic knowledge and classifications. Such classificatory building occurs within GSD spaces all over the globe. Eventually, these GSDs are brought into the Catalogue, which creates a consensus-based management classification to collocate and organize this data for global communication and sharing. As was discussed in chapter five, the final stage of Bourgoin's schematic gestures toward the use of the Catalogue—or any other biodiversity taxonomic platform—for large-scale evolutionary questions. This evolutionary reconciliation can only occur, however, among integrated taxonomies that can conform, or be mapped to, the taxonomic hierarchies and Linnaean nomenclature that have been prevalent in the Western production of science. We saw

how genetic barcodes are one source of integrative difficulty in this space, and how the Catalogue of Life Plus and GBIF were finding ways to address this epistemological and methodological gap.

But even despite the productive (and, indeed, necessary) increased centralization of biodiversity data, much work has yet to be done to facilitate the inclusion of data sources that fall outside of western epistemological traditions.[117] The accumulation of "normal" scientific progress articulated by Thomas Kuhn (1996) does not easily embrace knowledge produced outside of those established empirical traditions. For example, many scientists have articulated the importance of indigenous knowledge sets to the study of biodiversity issues (Mauro & Hardison, 2000), going so far as to call out the ethical responsibilities societies have to these marginalized natural history observations. Some projects have found methodologically sound ways to map indigenous knowledge sets with traditional, quantitative scientific forms. One such study by Clarence Alexander, et. al. (2011), found a correspondence between indigenous oral history narratives and scientific data based on geographic location information inherent in these data sources, and were thus able to present more robust climate maps depicting temperature change over time. Further, rich botanical and zoological manuscripts like the Mesoamerican Florentine Codex (Sahagun, 2012), rich in graphical and textual forms, chronicle natural history as conceptualized by historical communities. What of this knowledge in relation to our Linnaean based taxonomic systems? Even as integrative research of this nature proceeds, however, these data sources have yet to find their way into the data systems that are used to dictate climate policy and conservation around the globe. Indigenous knowledge has much to offer the general Western epistemology of biodiversity science (Agrawal, 1995; Cardoso, de Queiroz, Bandeira, &

[117] Certainly, this is *not* a critique on the iLife biodiversity landscape or the Catalogue of Life—collocating and organizing global data for one scientific tradition is difficult enough!

Góes-Neto, 2010). My next goal is to begin to theorize and practically address this gap by building upon the work performed in this dissertation.

My subsequent project will focus on immediately establishing an *Alternative-Synthetic-Taxonomy Laboratory* (*Alt-Syn-Tax*) that will bring together expertise from multiple domains (information communities, science & technology studies, and indigeneity studies) to develop systems that allow for the application of diverse classifications (and ontologies) using networked technologies, text-mining methodologies, and qualitative modes of text analysis. This research stream will make significant contributions to Information Studies by articulating the need for two disciplinary focuses: *Information and Documentary Diversity Studies* and *Interdisciplinary Knowledge Organization Studies*. Such a project can also inform the practices of biodiversity studies and biodiversity informatics. Information and Documentary Diversity Studies articulates ways in which traditional information and documentary approaches can be conceptualized as encompassing, and being augmented by, non-traditional documentary and descriptive forms. Interdisciplinary KO studies will focus on a more broadly defined boundaries for KO that allow for the inclusion and examination of new ontological and epistemological approaches. For biodiversity studies, it provides a mechanism through which contemporary scientific knowledge can communicate with a historically distinct cache of rich ethnohistorical and observational data in some productive way. Secondly, this project assumes that the epistemological rifts between Western science and other modes of understanding the natural world are *reconcilable*, at least to some degree, within *hermeneutic* taxonomic spaces.

I want to push the concept of retrieval-oriented modes of organization to its limit, and test the extensive power of systems like the Catalogue. As this dissertation has highlighted, all database technologies are interpretive spaces to some extent—they structure data in ways that

influence how we understand our historical practices (Bowker, 2008). This project acknowledges the role classification technologies plays in the *construction* of knowledge, and seeks to find ways to reconcile the epistemological boundaries that are part-and-parcel of these technical spaces. Finally, this project will bridge information theory with *practice* by developing more culturally inclusive information systems, and in the process of doing so, will engage underrepresented and counter-traditional community-generated knowledge sets into systems that impact biodiversity global policy. I want *Alt-Syn-Tax* to embrace uncertain spaces and create solutions for complex intellectual problems in a collaborative environment.

A core part of this endeavor is to (a) actively engage in global biodiversity policy domains so that governmental conservation practices and legal regimes represent diverse cultural concerns, (b) partner with museums and other cultural institutions to implement systems in practical settings, and (c) clearly articulate the professional and ethical responsibilities of Information Studies to pursue such diverse and extra-disciplinary endeavors. If the Global Taxonomic Initiative is any indication, there is a direct connection between *policy* and scientific practice. The classification systems we build (and the technologies that convey them) have a direct influence on how we conserve species, and how we understand and conceptualize climate and geographic issues of global concern (Edwards, 2010). Interviews I conducted for this study hinted at the complex policy regimes influenced directly by taxonomy. For one, as systems like ITIS create species lists, if one species is declared "invasive," the import and export of that species—and the funding model that arises from that exchange—is directly impacted. National income is, at times, dependent on how we categorize biological objects. Thus, there is a direct connection between information and documentation studies and policy, and this is an area where biodiversity can be useful in highlighting these connections.

Lastly, a more in-depth question about ontology is in order. If time permitted, I would have included a chapter on species ontology in this dissertation. More specifically, what *kinds* of characters and values constitute the description of a species—what kinds of questions are asked and what methods are invoked to *describe* and *identify* types of specimens as representative and constitutive of a natural object? How do scientists *compose* documentary evidence? And how do the processes and assumptions engrained in such descriptive practices differ from the application of bibliographic and documentary subject analysis?

Conclusion

Though the control of taxonomic data in knowledge databases is an acknowledged necessity in the biodiversity landscape, the question remains of how we define that control in relation to the shifting knowledge, object, and subject conditions that define the practices of scientists, and the work, hypothesizing, and paradigms of scientific communities. In Patrick Wilson's concluding chapter to *Two Kinds of Power* (1968) he proposes a hypothetical Supreme Bibliographic Council,

> Whose task it was to evaluate the bibliographical situations of individuals and groups of individuals, to estimate the degree of bibliographical control available to them, to decide on its adequacy or inadequacy, and to suggest or order changes in those situations, by the creation of new bibliographical instruments or new institutional arrangements, or by the alteration of old arrangements, or by making more widely available instruments and services hitherto restricted in availability (1968, p. 132).

So much of the power possible in Wilson's bibliographical universe is contingent in the sense that instrumentalized control exerts only as much power as is applicable in a given context and known to a given individual. That a Supreme Council is even postulated as an intellectual exercise emphasizes the fact that a documentary instrument is, by definition, never complete, and perhaps more importantly, always *potentially* deficient in certain circumstance. Even if all extant documentation could be aggregated into one system (impossible though this may be), that aggregation (and its attendant control mechanisms and imposed relationship structure)

presupposes a certain set of uses, and such presuppositions will inevitably lead to information deficiencies for *some* set of individuals. Instrumental need is always idiosyncratic and contextually specific, while an instrument's architecture must embody the ability to embody and attend to diverse approaches, multiple conditions, and exemplifications. It is this tension that lies at the heart of this manuscript's discourse.

Knowledge organization systems *contain* information—not only as containers contain a set of things or objects for aesthetic or functional convenience, but as containers *inhibit* the movement of said items within the bounds of its edges (or, as Hope Olson, might say, *limits*) (Olson, 2002). Systems construct and organize information into specific arrangements of knowledge. Having *control* in KO systems is equivalent to making an epistemological, ontological, and interpretive commitment to a kind of subjective knowledge. And such contours have direct influence on what kinds of powers people can have in relation to that system. The question becomes, how much control must you give up to make an optimally flexible KO system? We see this conflict playing-out between descriptive-oriented and retrieval-oriented systems in the biodiversity community. The two, perhaps, shall never meet at a workable medium. One community understands classifications as hermeneutic spaces, intent on modeling an accurate representation of the natural world, while the other sees these instruments as mediating, switching, data control spaces that facilitate a broader understanding of taxonomic practice as a object form in-and-of itself. In the biodiversity documentation environment—as is the case in Wilson's bibliographical universe—the notion of access is not simply the satisfaction of a specific *need* for satisfaction's sake, for no person can possibly know what they do not know. Systems must also be used to locate what was otherwise unsought.

Classification instrumentation must present combinatory and flexible knowledge spaces so as to produce new forms of knowledge, both intended and unintended. A Supreme Council will not and cannot exist to mediate access, so our systems must control for this usage contingency. The documents of biodiversity knowledge are far more than the species concepts taxonomies intend to organize, they are also a constellation of evidence, all of which *will be* used to recursively and consistently reevaluate taxonomic conclusions to produce new hypotheses. What we see in biodiversity spaces at the moment is a deep attention to mediating the expectations between those that want taxonomy to primarily *describe*, versus those that see the need for taxonomy to facilitate information *retrieval*.

The larger ecological, political, and social consequences of taxonomies are clearly on the minds of those working for the Catalogue and in the iLife consortium. Their biodiversity scientific and taxonomic work is in service to the broader goals of conservation and a deep appreciation for the species they dedicate their lives to studying. Taxonomic and classificatory work is facilitative: it is a service provided so that *some kind* of action can then be taken toward some social, ecological, or political goal. As Patrick Wilson states,

> Findings on adequacy and inadequacy, on the part of a Supreme Bibliographic Council, will inevitably be political decisions … whether the policies that guide its decisions are stated in terms of exploitative or descriptive control or both, its decisions about who shall have how much of what sort of control will be decisions perfectly parallel in character to the decisions of other political-decision making bodies…There is a scientific and literary patrimony, as Langlois calls it, to be managed and exploited; men must decide whether they are content to record the existence of writings and store them up in repositories, or whether they wish to pursue the active exploitation of the patrimony, to provide means of making the maximum use of the useable writings. We can, by reflection and by experimentation, make clear the possible goals and discover and text devices for the attainment of those goals; it is only by a political decision that one goal can be singled out as the "proper" goal, that it can be said who is to have how much of the power over writings, and the knowledge contained in them, that bibliographical control confers (1968, p. 155).

The Catalogue is one primary way by which the biodiversity community is attempting to gain an aggregative control over the global production of taxonomic knowledge, setting the stage for the production of new kinds of biodiversity knowledge. Hope Olson, too, called for *eccentric*

techniques (2002) in order to break free from the limitations imposed by the cultural and social presuppositions that shape the categories we identify in our world.

So, are consensus-based taxonomies, then, an implementable method by which we can conceptualize more traditional documentary and bibliographic classificatory systems? I suppose all I can say at this point is that the jury is still out. But despite some stated uncertainty as to the Catalogue's general *current* success, there is a sense that it is undoubtedly proceeding in the correct direction. In a conversation with a prominent and globally renowned taxonomist, it was stated that infrastructure like the Catalogue could be a game changer for on-the-ground scientific work: "Oh god, people like me [are the audience for the Catalogue]. If I could trust those things, god I would use them all the time. [But], you [need to] get a bunch of obsessive compulsive people, clone them, and get them to sit down with these data sets and clean them up … You want there to be that validation…." *Validation* and *trust* is an essential part of this taxonomic space. Scientists are aware of the limitations of the systems *they* create, so they see their limitations, but judging the efficacy of *other* systems is a far more difficult prospect. Until—or, perhaps, *when*—systems like the Catalogue can document their process fully enough to garner general trust, then the CERN-like infrastructure biodiversity scientists are striving for will come to fruition. The end result, of course, which is of central concern for my scholarly agenda, is that we better understand classification systems so that they can be more just. They must *both* fairly represent the communities and objects they intend to document, and be equally distributed to and representative of these communities as a form of public knowledge. My goal here has been to show how other disciplines, such as biodiversity studies, can inform the literature, theories, and work within Information Studies to rethink our problems anew. As Wilson indicates above, experimentation is key, lest our systems fail to provide the one service they set out to offer:

situational relevant information and the enhancement of intellectual, social, and knowledge-based power. As I see it, the question is no longer whether our classification systems can attain true representational capacities, it is more about how we are going to acknowledge their constructedness and harness their contingencies for the most social and community-driven benefit.

Appendix

Appendix A

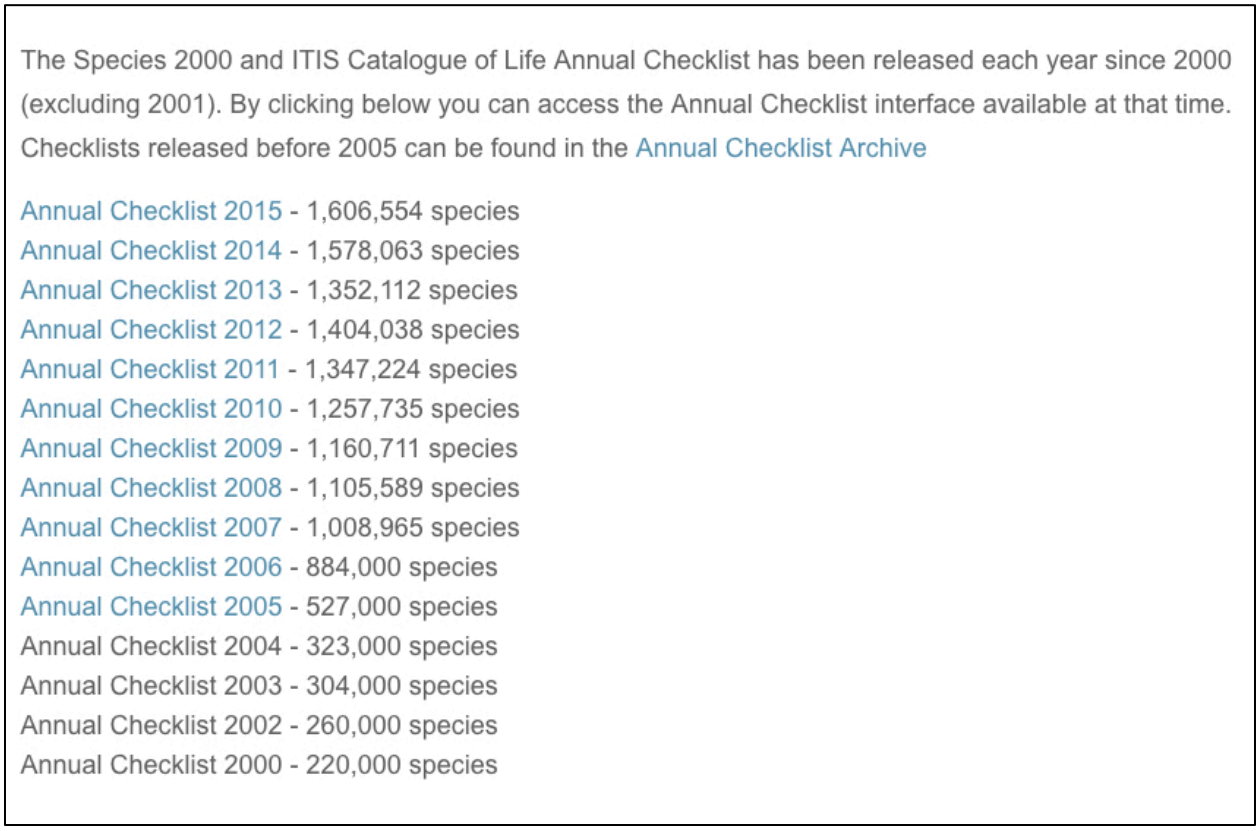


Figure 27. Species Counts in the Catalogue of Life for the years 2000-2015 (Species 2000, 2017d).

References

Adl, S. M., Simpson, A. G. B., Farmer, M. A., Andersen, R. A., Anderson, O. R., Barta, J. R., … Taylor, M. F. J. R. (2005). The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *The Journal of Eukaryotic Microbiology*, *52*(5), 399–451. https://doi.org/10.1111/j.1550-7408.2005.00053.x

Adler, M., Tennis, J. T., Martinez-Ávila, D., Guimarães, J. A. C., Mai, J., Olesen-Bagneux, O., & Skouvig, L. (2016). Global/local knowledge organization: Contexts and questions. *Proceedings of the 79th ASIS&T Annual Meeting*, *53*.

Agrawal, A. (1995). Indigenous and scientific knowledge: Some critical comments. *Indigenous Knowledge and Development Monitor*, *3*(3).

Alexander, C., Bynum, N., Johnson, E., King, U., Mustonen, T., Neofotis, P., … Weeks, B. (2011). Linking Indigenous and scientific knowledge of climate change. *BioScience*, *61*(6), 477–484. https://doi.org/10.1525/bio.2011.61.6.10

Allen, D., & Ellis, D. (2000). The paradigm debate in information systems research. *Information Systems Review*, (1), 21–28.

Anderson, R., Araújo, M., Guisan, A., Lobo, J. M., Martínez-Meyer, E., Peterson, A. T., & Soberón, J. (2016, March 22). Report of the task group on GBIF data fitness for use in distribution modelling [Text]. Retrieved February 28, 2017, from http://www.gbif.org/resource/82612

Asundi, A. Y. (2012). Domain specific categories and relations and their potential applications: A case study of two arrays of agriculture schedule of colon classification. *Advances in Knowledge Organization*, *13*(28).

Atkin, A. (2013). Peirce's Theory of Signs. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2013). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/sum2013/entries/peirce-semiotics/

Atlas of Living Australia. (2016). Retrieved February 10, 2017, from http://www.ala.org.au/

Baron, M. G., Norman, D. B., & Barrett, P. M. (2017). A new hypothesis of dinosaur relationships and early dinosaur evolution. *Nature*, *543*(7646), 501–506. https://doi.org/10.1038/nature21700

Bastian, J. A. (2002). Taking custody, giving access: a postcustodial role for a new century. *Archivaria*, *1*(53).

Bates, M. J. (1999). The invisible substrate of information. *Journal of the American Society for Information Science*, *50*(12), 1043–1050.

Bates, M. J. (2006). Fundamental forms of information. *Journal of the American Society for Information Science and Technology*, *57*(8), 1033–1045. https://doi.org/10.1002/asi.20369

Bates, M. J. (2007). Defining the information disciplines in encyclopedia development. In *Proceedings of the Sixth International Conference on Conceptions of Library and Information Science* (Vol. 12:4). Borås, Sweden: Information Research. Retrieved from http://www.informationr.net/ir/12-4/colis/colis29.html

Bates, M. J. (2009). Information. In *Encyclopedia of Library and Information Sciences, Third Edition* (pp. 2347–2360). Taylor & Francis. Retrieved from http://www.tandfonline.com/doi/abs/10.1081/E-ELIS3-120045519

Beghtol, C. (1986). Semantic validity: concepts of warrant in bibliographic classification systems. *Library Resources & Technical Services*, *30*, 109–25.

Beghtol, C., Green, R., & Bean, C. A. (2001). Relationships in classificatory structures and meaning. In *Relationships in the organization of knowledge* (pp. 99–113). Dordrecht ; Boston : Norwell, MA: Kluwer Academic Publishers ; Sold and distributed in North, Central, and S. America by Kluwer Academic Publishers.

Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., & Courchamp, F. (2012). Impacts of climate change on the future of biodiversity. *Ecology Letters*, *15*(4), 365–377. https://doi.org/10.1111/j.1461-0248.2011.01736.x

Berendsohn, W. G. (1995). The concept of "potential taxa" in databases. *Taxon*, *44*(2), 207–212. https://doi.org/10.2307/1222443

Berendsohn, W. G., & Geoffrey, M. (2007). Networking concepts: Uniting without "Unitary-ism." In C. J. Humphries & G. B. Curry (Eds.), *Biodiversity databases: techniques, politics, and applications*. Boca Raton: CRC Press.

Bernd Frohmann. (2009). Revisiting "What is a document?" *Journal of Documentation*, *65*(2), 291–303. https://doi.org/10.1108/00220410910937624

Biodiversity Data Journal. (2017). Biodiversity Data Journal. Retrieved March 4, 2017, from http://bdj.pensoft.net/articles.php?id=

Biodiversity Heritage Library. (2017). Biodiversity Heritage Library - About. Retrieved March 20, 2017, from http://biodivlib.wikispaces.com/About

Bisby, F. A. (2000). The quiet revolution: Biodiversity informatics and the internet. *Science*, *289*(5488), 2309–2312. https://doi.org/10.1126/science.289.5488.2309

Bisby, F. A., Shimura, J., Ruggiero, M., Edwards, J., & Haeuser, C. (2002). Taxonomy, at the click of a mouse. *Nature*, *418*(6896), 367–367. https://doi.org/10.1038/418367a

Blair, D. C. (1984). The data-document distinction in information retrieval. *Commun. ACM*, *27*(4), 369–374. https://doi.org/10.1145/358027.358049

Blair, D. C. (2006). The data-document distinction revisited. *SIGMIS Database*, *37*(1), 77–96. https://doi.org/10.1145/1120501.1120507

Blake, J. (2011). Some Issues in the classification of zoology. *Knowledge Organization*, *38*(6), 463–472.

Blanchette, J.-F. (2011). A material history of bits. *Journal of the American Society for Information Science and Technology*, *62*(6), 1042–1057. https://doi.org/10.1002/asi.21542

Blaxter, M. L. (2004). The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *359*(1444), 669–679. https://doi.org/10.1098/rstb.2003.1447

Bliss, H. E. (1929). *The organization of knowledge and the system of the sciences*. New York: H. Holt and Company. Retrieved from http://catalog.hathitrust.org/Record/001388383

Bliss, H. E. (1933). *The organization of knowledge in libraries and the subject-approach to books*. New York , NY: The H.W. Wilson Company.

Borgman, C. L. (2015). *Big data, little data, no data: scholarship in the networked world*. Cambridge, Massachusetts: The MIT Press.

Bourgoin, T. (2016, November). *Importance of databasing taxonomic knowledge: an example with FLOW (Fulgoromorpha Lists on the Web). (Invited Communication)*. Beijing Forestry University, Beijing, China.

Bowers, F. (1994). *Principles of bibliographical description*. Winchester, U.K. : New Castle, DE: St. Paul's Bibliographies ; Oak Knoll Press.

Bowker, G. C. (2000a). Biodiversity datadiversity. *Social Studies of Science*, *30*(5), 643–683. https://doi.org/10.1177/030631200030005001

Bowker, G. C. (2000b). Mapping biodiversity. *International Journal of Geographical Information Science*, *14*(8), 739–754. https://doi.org/10.1080/136588100750022769

Bowker, G. C. (2008). *Memory practices in the sciences*. Cambridge, Mass.; London: The MIT Press.

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. Cambridge, Mass: MIT Press.

Bowker, G. C., Timmermans, S., Clarke, A. E., & Balka, E. (Eds.). (2015). *Boundary objects and beyond: Working with Leigh Star*. Cambridge, MA: The MIT Press.

Boyle, B., Hopkins, N., Lu, Z., Raygoza Garay, J. A., Mozzherin, D., Rees, T., … Enquist, B. J. (2013). The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics*, *14*, 16. https://doi.org/10.1186/1471-2105-14-16

Briet, S. (1951). *What is documentation? English translation of the classic French text*. (R. E. Day, Trans.). Lanham, Md: Scarecrow Press.

Broadfield, A. (1946). *The philosophy of classification* (1st ed.). London: Grafton & Co.

Buchanan, B. (1979). *Theory of library classification*. London : New York: C. Bingley ; K.G. Saur.

Buckland, M. (1991). Information as thing. *Journal of the American Society for Information Science*, *42*(5), 351.

Buckland, M. (1997). What is a "document"? *Journal of the American Society for Information Science*, *48*(9), 804–809.

Buckland, M. (1998). What is a "digital document"? *Journal of the American Society for Information Science*, *48*(9), 215–220.

Butchart, S. H. M., Walpole, M., Collen, B., Strien, A. van, Scharlemann, J. P. W., Almond, R. E. A., … Watson, R. (2010). Global biodiversity: Indicators of recent declines. *Science*, *328*(5982), 1164–1168. https://doi.org/10.1126/science.1187512

Cachuela-Palacio, M. (2006). Towards an index of all known species: the Catalogue of Life, its rationale, design and use. *Integrative Zoology*, *1*(1), 18–21. https://doi.org/10.1111/j.1749-4877.2006.00007.x

Capurro, R., & Hjørland, B. (2003). The concept of information. *Annual Review of Information Science and Technology*, *37*(1), 343–411. https://doi.org/10.1002/aris.1440370109

Cardoso, D. B. O. S., de Queiroz, L. P., Bandeira, F. P., & Góes-Neto, A. (2010). Correlations between indigenous Brazilian folk classifications of fungi and their systematics. *Journal of Ethnobiology*, *30*(2), 252–264. https://doi.org/10.2993/0278-0771-30.2.252

Carlyle, A. (2006). Understanding FRBR as a conceptual model: FRBR and the bibliographic universe. *Library Resources & Technical Services*, *50*(4), 264–273.

Caswell, M. (2014). Toward a survivor-centered approach to records documenting human rights abuse: lessons from community archives. *Archival Science*, *14*(3–4), 307–322. https://doi.org/10.1007/s10502-014-9220-6

Cavalier-Smith, T. (1998). A revised six-kingdom system of life. *Biological Reviews*, *73*(3), 203–266. https://doi.org/10.1111/j.1469-185X.1998.tb00030.x

CKAN. (2017). ckan. Retrieved April 2, 2017, from https://ckan.org/

Claudio Gnoli, & Riccardo Ridi. (2014). Unified theory of information, hypertextuality and levels of reality. *Journal of Documentation*, *70*(3), 443–460. https://doi.org/10.1108/JD-09-2012-0115

composite, adj. and n. (2016). *OED Online*. Oxford University Press. Retrieved from http://www.oed.com/view/Entry/37791

Convention on Biological Diversity. (2003). Guide to the Global Taxonomic Initiative. Secretariat of the Convention on Biological Diversity. Retrieved from https://www.cbd.int/doc/publications/cbd-ts-30.pdf

Convention on Biological Diversity. (2016). History of the Convention. Retrieved from https://www.cbd.int/history/default.shtml

Convention on Biological Diversity. (2017a). Conference of the Parties (COP). Retrieved March 18, 2017, from https://www.cbd.int/cop/

Convention on Biological Diversity. (2017b). Global Taxonomy Initiative: Background. Retrieved from https://www.cbd.int/gti/background.shtml

Convention on Biological Diversity. (2017c). The Convention on Biological Diversity [cbd.org]. Retrieved March 18, 2017, from https://www.cbd.int/convention/

Convention on Biological Diversity (full text). (1992). United Nations. Retrieved from https://www.cbd.int/doc/legal/cbd-en.pdf

Cook, T. (1994). Electronic records, paper minds: the revolution in information management and archives in the post/ custodial and post/ modernist era. *Archives and Manuscripts*, *22*(2).

Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval*, *7*(1), 19–37. https://doi.org/10.1016/0020-0271(71)90024-6

Coyle, K. (2016). *FRBR, before and after: a look at our bibliographic models*. Chicago: ALA Editions, an imprint of the American Library Association.

Coyle, K., & Hillmann, D. (2007). Resource Description and Access (RDA): Cataloging Rules for the 20th Century. *D-Lib Magazine*, *13*(1/2). https://doi.org/10.1045/january2007-coyle

Croft, J., Cross, N., Hinchcliffe, S., Lughadha, E. N., Stevens, P. F., West, J. G., & Whitbread, G. (1999). Plant Names for the 21st Century: The International Plant Names Index, a Distributed Data Source of General Accessibility. *Taxon*, *48*(2), 317–324. https://doi.org/10.2307/1224436

Cronin, B. (2008). The sociological turn in information science. *Journal of Information Science*, *34*(4), 465–475. https://doi.org/10.1177/0165551508088944

Current status of the Myriapod Class Diplopoda (Millipedes): Taxonomic diversity and phylogeny. (2007). *Annual Review of Entomology*, *52*(1), 401–420. https://doi.org/10.1146/annurev.ento.52.111805.090210

Dane, J. A. (1995). Ideal copy "versus" ideal texts': The application of bibliographical description to facsimiles. *Papers of The Bibliographical Society of Canada*, *33.1*.

Darwin, C. (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, Albemarle Street.

Darwin Core Task Group. (2011, 2015). Darwin Core. Retrieved February 26, 2017, from http://rs.tdwg.org/dwc/

Daston, L. (2004). Type specimens and scientific memory. *Critical Inquiry*, *31*(1), 153–182. https://doi.org/10.1086/427306

Daston, L., & Galison, P. (2007). *Objectivity*. New York; Cambridge, Mass.: Zone Books ; Distributed by the MIT Press.

Daston, L., & Lunbeck, E. (Eds.). (2011). *Histories of scientific observation*. Chicago ; London: University Of Chicago Press.

Day, R. E. (2008). *The modern invention of Information discourse, history, and power*. Carbondale: Southern Illinois University Press. Retrieved from http://site.ebrary.com/id/10695251

Day, R. E. (2014). *Indexing it all: The subject in the age of documentation, information, and data* (1 edition). Cambridge, Massachusetts: The MIT Press.

Dayrat, B. (2005). Towards integrative taxonomy. *Biological Journal of the Linnean Society*, *85*(3), 407–415. https://doi.org/10.1111/j.1095-8312.2005.00503.x

De Queiroz, K. (2005). A unified concept of species and its consequences for the future of taxonomy. *Proceedings from the California Academy of Sciences*, *56*(18), 196–215.

Dekker, R. (2016, September 7). Personal Interview.

Deleuze, G. (1987). *A thousand plateaus: capitalism and schizophrenia*. Minneapolis: University of Minnesota Press.

Denton, W., & Taylor, A. G. (Eds.). (2007). FRBR and the History of Cataloging. In *Understanding FRBR: what it is and how it will affect our retrieval tools*. Westport, Conn: Libraries Unlimited.

Deokattey, S., Neelameghan, A., & Kumar, V. (2010). A method for developing a domain ontology: A case study for a multidisciplinary subject. *Knowledge Organization*, *37*(3), 173–184.

Department of Public Information. (1997, May 23). UN conference on environment and development (1992). Retrieved from http://www.un.org/geninfo/bp/enviro.html

Dewey, S. H. (2014). The Continuing relevance of Paul Otlet, the International Institute of Bibliography/International Federation for Documentation, and the Documentation Movement for Information Science and Studies. *InterActions: UCLA Journal of Education and Information Studies*, *10*(2). Retrieved from http://escholarship.org/uc/item/5pq3v1cp

Döring, E. von M. (2016, April 6). Updating the GBIF backbone [Blog.]. Retrieved December 21, 2016, from http://gbif.blogspot.com/2016/04/updating-gbif-backbone.html

Dornelas, M., Gotelli, N. J., McGill, B., Shimadzu, H., Moyes, F., Sievers, C., & Magurran, A. E. (2014). Assemblage time series reveal biodiversity change but not systematic loss. *Science*, *344*(6181), 296–299. https://doi.org/10.1126/science.1248484

Drucker, J. (2013). *What Is?: Nine Epistemological Essays* (First edition). Victoria, Texas: Cuneiform Press.

Drucker, J. (2014). Distributed and conditional documents: conceptualizing bibliographical alterities. *Materialities of Literature*, *2*(1), 11–29. https://doi.org/10.14195/2182-8830_2-1_1

Dupré, J. (1993). *The disorder of things: metaphysical foundations of the disunity of science*. Cambridge, Mass: Harvard University Press.

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. Cambridge, Mass: MIT Press.

Eldredge, N., & Cracraft, J. (1980). *Phylogenetic patterns and the evolutionary process: method and theory in comparative biology*. New York: Columbia University Press.

Ellis, D. (1992). Paradigms and proto-paradigms in information retrieval research. In *Conceptions of Library and Information Science. Historical, empirical and theoretical perspectives* (pp. 165–186). London: T. Graham.

Encyclopedia of Life. (2016). EoL.org. Retrieved November 1, 2016, from http://www.eol.org/

Encyclopedia of Life. (2017a). Brown Bear, Grizzly Bear - Ursus arctos - All Classifications. Retrieved February 12, 2017, from http://eol.org/pages/328581/names

Encyclopedia of Life. (2017b). Brown Bear, Grizzly Bear - Ursus arctos - Classifications. Retrieved February 12, 2017, from http://eol.org/pages/328581/names

Encyclopedia of Life. (2017c). Brown Bear, Grizzly Bear - Ursus arctos - Overview. Retrieved February 12, 2017, from http://eol.org/pages/328581/overview

Encyclopedia of Life. (2017d). Classification Providers. Retrieved February 12, 2017, from http://eol.org/info/222

Encyclopedia of Life: Global access to knowledge about life on Earth. (2015). Retrieved October 1, 2015, from http://eol.org/

Eng, K. (2016, April 21). A newly drawn Tree of Life reminds us to question what we know. Retrieved December 24, 2016, from http://ideas.ted.com/a-newly-drawn-tree-of-life-reminds-us-to-question-what-we-know/

EOL Curators. (2016). Retrieved February 12, 2017, from http://160.111.248.29/curators

Ereshefsky, M. (2001). Names, numbers and indentations: a guide to post-Linnaean taxonomy. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *32*(2), 361–383. https://doi.org/10.1016/S1369-8486(01)00004-8

Ereshefsky, M. (2007). *The poverty of the Linnaean hierarchy: A philosophical study of biological taxonomy* (1 edition). Cambridge England: Cambridge University Press.

Erickson, D. L., & Driskell, A. C. (2012). Construction and analysis of phylogenetic trees Using DNA barcode data. In W. J. Kress & D. L. Erickson (Eds.), *DNA Barcodes* (Vol. 858, pp. 395–408). Totowa, NJ: Humana Press. Retrieved from http://link.springer.com/10.1007/978-1-61779-591-6_19

European Organization for Nuclear Research. (2017). CERN | Accelerating science. Retrieved March 17, 2017, from https://home.cern/

Evans, C. (2013, November 20). University of Reading scientists hand over unique list of more than 1.4 million species to Dutch Research Centre - Get Reading. Retrieved July 16, 2015, from http://www.getreading.co.uk/news/local-news/university-reading-scientists-hand-over-6323483

Federhen, S. (2003). The Taxonomy Project. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK21100/

Fishbase. (2017, February). FishBase. Retrieved April 1, 2017, from http://www.fishbase.org/search.php

Floridi, L. (Ed.). (2004). *The Blackwell guide to the philosophy of computing and information*. Malden, MA: Blackwell Pub.

Floridi, L. (2010). *Information: a very short introduction*. Oxford ; New York: Oxford University Press.

Floridi, L. (2011). *The philosophy of information* (1. publ. in paperback). Oxford: Oxford Univ. Press.

Foucault, M. (1994). *The order of things: an archaeology of the human sciences*. New York: Vintage Books.

Franz, N. M. (2005). On the lack of good scientific reasons for the growing phylogeny/classification gap. *Cladistics*, *21*(5), 495–500. https://doi.org/10.1111/j.1096-0031.2005.00080.x

Franz, N. M., Chen, M., Kianmajd, P., Yu, S., Bowers, S., Weakley, A. S., & Ludäscher, B. (2016). Names are not good enough: Reasoning over taxonomic change in the Andropogon complex1. *Semantic Web*, *7*(6), 645–667. https://doi.org/10.3233/SW-160220

Franz, N., Peet, R. K., & Weakley, A. S. (2008). On the use of taxonomic concepts in support of biodiversity research and taxonomy. In Q. Wheeler (Ed.), *The New taxonomy*. Boca Raton: CRC Press.

Franz [@taxonbytes], N. (2016, August 17). @timrobertson100 @GBIF @rdmpage @aeolid Yes, thank you. But that means there are no primary authors? No people to directly dis-/agree with. [microblog]. Retrieved March 4, 2017, from [ADD URL FROM: https://storify.com/taxonbytes/who-authors-gbif-s-backbone; listed but on plane, could not locate]

Furner, J. (2004a). Conceptual analysis: A method for understanding information as evidence, and evidence as information. *Archival Science*, *4*(3–4), 233–265. https://doi.org/10.1007/s10502-005-2594-8

Furner, J. (2004b). Information studies without information. Retrieved from https://www.ideals.illinois.edu/handle/2142/1684

Furner, J. (2009). Interrogating identity : A philosophical approach to an enduring issue in knowledge organization. *Knowledge Organization*, *36*(1), 3–16.

Furner, J. (2013a). ASIS&T annual meeting pre-conference activities: The 23rd annual SIG/CR classification research workshop: A report. *Bulletin of the American Society for Information Science and Technology*, *39*(3), 28–32. https://doi.org/10.1002/bult.2013.1720390309

Furner, J. (2013b, Fall). *What is Information?* Classroom Lecture presented at the Doctoral Seminar: Theoretical Traditions (IS291A), University of California, Los Angeles.

Furner, J. (2014). Information without information studies. In *Theories of Information, Communication and Knowledge*. Netherlands: Springer Netherlands.

Furner, J. (2015). Information science is neither. *Library Trends*, *63*(3), 362–377. https://doi.org/10.1353/lib.2015.0009

Furner, J. (2016a). "Data": The data. In M. Kelly & J. Bielby (Eds.), *Information Cultures in the Digital Age* (pp. 287–306). Wiesbaden: Springer Fachmedien Wiesbaden. Retrieved from http://link.springer.com/10.1007/978-3-658-14681-8_17

Furner, J. (2016b). Type-token theory and bibliometrics. In C. R. Sugimoto & B. Cronin (Eds.), *Theories of informetrics and scholarly communication: a Festschrift in honor of Blaise Cronin*. Berlin: De Gruyter.

Galison, P., & Hevly, B. W. (Eds.). (1992). *Big science: the growth of large-scale research*. Stanford, Calif: Stanford University Press.

GBIF. (2011). GBIF name parser [Text]. Retrieved January 11, 2017, from http://www.gbif.org/resource/81521

GBIF. (2013, August 19). The GBIF Secretariat [Text]. Retrieved April 1, 2017, from http://www.gbif.org/governance/secretariat

GBIF. (2016a). Occurrence records. Retrieved February 26, 2017, from http://www.gbif.org/occurrence

GBIF. (2016b, September 21). Ecuador joins GBIF as an associate participant [Text]. Retrieved April 1, 2017, from http://www.gbif.org/newsroom/news/ecuador-joins-gbif

GBIF. (2017). GBIF/name-parser. Retrieved January 11, 2017, from https://github.com/gbif/name-parser

GBIF. (n.d.). Participant list [Text]. Retrieved April 1, 2017, from http://www.gbif.org/participation/participant-list

GBIF backbone taxonomy - Constituents. (2017). Retrieved December 24, 2016, from http://www.gbif.org/dataset/d7dddbf4-2cf0-4f39-9b2a-bb099caae36c/constituents

GBIF Science Committee. (2016, October 10). GBIF science review 2016. (R. D. M. Page, M. Costello, A. G. Finstad, P. Grandcolas, E. Arnaud, & G. Cochrane, Eds.). Retrieved from http://www.gbif.org/resource/82873

gbif/checklistbank schema. (2017). Retrieved March 4, 2017, from https://github.com/gbif/checklistbank/blob/master/docs/schema.pdf

GBIF.org. (2016, February 25). Data processing [Text]. Retrieved March 1, 2017, from http://www.gbif.org/infrastructure/processing

GBIF.org. (2017). Occurrence detail 236047012. Retrieved March 1, 2017, from http://www.gbif.org/occurrence/236047012#issues

Gewin, V. (2002). Taxonomy: All living things, online. *Nature*, *418*(6896), 362–363. https://doi.org/10.1038/418362a

Ghosh, P. (2017, March 22). Major shake-up suggests dinosaurs may have "UK origin." *BBC News*. Retrieved from http://www.bbc.com/news/science-environment-39305750

Gitelman, L. (Ed.). (2013). *"Raw Data" Is an Oxymoron*. Cambridge, Massachusetts ; London, England: The MIT Press.

Global Biodiversity and Information Facility. (2013a, August 19). Publishing data [Text]. Retrieved March 4, 2017, from http://www.gbif.org/publishing-data/summary

Global Biodiversity and Information Facility. (2013b, August 19). What is GBIF? [Text]. Retrieved March 4, 2017, from http://www.gbif.org/what-is-gbif

Global Biodiversity and Information Facility. (2015). GBIF. Retrieved October 1, 2015, from http://www.gbif.org/

Global Biodiversity and Information Facility. (2016a). GBIF backbone taxonomy. Retrieved July 15, 2016, from http://www.gbif.org/dataset/d7dddbf4-2cf0-4f39-9b2a-bb099caae36c/stats

Global Biodiversity and Information Facility. (2016b, December 15). "Names in November" workshop targets single shared species list [Text]. Retrieved January 21, 2017, from http://www.gbif.org/newsroom/news/species-names-in-november-workshop

Global Biodiversity and Information Facility. (2017a). gbif/checklistbank. Retrieved March 4, 2017, from https://github.com/gbif/checklistbank

Global Biodiversity and Information Facility. (2017b). mdoering/backbone. Retrieved March 4, 2017, from https://github.com/mdoering/backbone

Global Names Architecture. (2015, September 23). Global Names Usage Bank (GNUB). Retrieved January 21, 2017, from http://globalnames.org/docs/gnub/

Global Names Architecture. (2016a). Good and bad names. Retrieved January 20, 2017, from http://globalnames.org/docs/good-bad-names/

Global Names Architecture. (2016b, November 15). Names in November meeting. Retrieved January 21, 2017, from http://globalnames.org/news/2016/11/15/names-in-november/

Global Names Architecture. (2017a). Global Names recognition and discovery tools and services. Retrieved April 10, 2017, from http://gnrd.globalnames.org/

Global Names Architecture. (2017b). Global Names Resolver. Retrieved January 20, 2017, from http://resolver.globalnames.org/

Global Names Architecture. (2017c). GlobalNames Home. Retrieved January 21, 2017, from http://globalnames.org/

Global Names Architecture. (2017d). Global Names Architecture: Glossary. Retrieved January 11, 2017, from http://globalnames.org/docs/glossary/

Gnoli, C. (2006). Phylogenetic classification. *Knowledge Organization*, *33*(3), 138–152.

Gnoli, C. (2009, 2011). Integrative classification: A principle for knowledge organization. Retrieved February 11, 2017, from http://www.iskoi.org/ilc/book/principle.php

Gnoli, C., & Poli, R. (2004). Levels of reality and levels of representation. *Knowledge Organization*, *31*(3), 151–160.

Godfray, H. C. J. (2002). Challenges for taxonomy. *Nature*, *417*(6884), 17–19. https://doi.org/10.1038/417017a

Godfray, H. C. J. (2007). Linnaeus in the information age. *Nature*, *446*(7133), 259–260. https://doi.org/10.1038/446259a

Gordon, D. P. (2009). Towards a management hierarchy (classification) for the Catalogue of Life: Draft Discussion Document. Species 2000 & ITIS Catalogue of Life: 2009 Annual Checklist. Retrieved from http://www.catalogueoflife.org/col/info/hierarchy

Gorichanaz, T., & Latham, K. F. (2016). Document phenomenology: a framework for holistic analysis. *Journal of Documentation*, *72*(6), 1114–1133. https://doi.org/10.1108/JD-01-2016-0007

Grant, V. (2003). Incongruence between cladistic and taxonomic systems. *American Journal of Botany*, *90*(9), 1263–1270. https://doi.org/10.3732/ajb.90.9.1263

Green, R. (2008). Relationships in knowledge organization. *Knowledge Organization*, *35*(2/3), 150–159.

Green, R., & Martin, G. (2013). A rosid is a rosid is a rosid . . . or not. *Advances in Classification Research Online*, *23*(1), 9–16. https://doi.org/10.7152/acro.v23i1.14228

Guala, G. F. (2016). The Importance of Species Name Synonyms in Literature Searches. *PLOS ONE*, *11*(9), e0162648. https://doi.org/10.1371/journal.pone.0162648

Guralnick, R., & Hill, A. (2009). Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics*, *25*(4), 421–428. https://doi.org/10.1093/bioinformatics/btn659

Haeckel, E. (1866). *Generelle morphologie der organismen. Allgemeine grundzüge der organischen formen-wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte descendenztheorie*. Berlin: Druck und Verlag von Georg Reimer.

Hamer, M., Victor, J., & Smith, G. F. (2012). Best Practice Guide for Compiling, Maintaining and Disseminating National Species Checklists. Global Biodiversity Information Facility. Retrieved from http://www.gbif.org/orc/?doc_id=4752

He, Y., Caporaso, J. G., Jiang, X.-T., Sheng, H.-F., Huse, S. M., Rideout, J. R., … Zhou, H.-W. (2015). Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome*, *3*(1). https://doi.org/10.1186/s40168-015-0081-x

Hennig, W. D. (1950). *Grundzüge einer Theorie der phylogenetischen Systematik.* Berlin, Deutscher Zentralverlag,.

Hennig, W., Davis, D. D., & Zangerl, R. (1999). *Phylogenetic systematics*. Urbana: University of Illinois Press.

Henry, L. (1998). Schellenberg in Cyberspace. *The American Archivist*, *61*(2), 309–327. https://doi.org/10.17723/aarc.61.2.f493110467x38701

Hinchliff, C. E., Smith, S. A., Allman, J. F., Burleigh, J. G., Chaudhary, R., Coghill, L. M., … Cranston, K. A. (2015). Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences*, *112*(41), 12764–12769. https://doi.org/10.1073/pnas.1423041112

Hjørland, B. (1998). The classification of psychology: A case study in the classification of a knowledge field. *Knowledge Organization. International Journal Devoted to Concept Theory, Classification, Indexing and Knowledge Representation*, *24*(4), 162–201.

Hjørland, B. (2008). What is knowledge organization (KO)? *Knowledge Organization. International Journal Devoted to Concept Theory, Classification, Indexing and Knowledge Representation*. Retrieved from http://arizona.openrepository.com/arizona/handle/10150/106183

Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: Domain-analysis. *Journal of the American Society for Information Science*, *46*(6), 400–425. https://doi.org/10.1002/(SICI)1097-4571(199507)46:6<400::AID-ASI2>3.0.CO;2-Y

Hjørland, B., & Hartel, J. (2003). Afterword: Ontological, epistemological and sociological dimensions of domains. *Knowledge Organization*, *30*(3/4).

Hjørland, B., & Nicolaisen, J. (2003). Scientific and scholarly classifications are not "Naïve": A comment to Begthol. *Knowledge Organization*, *31*(1), 55–61.

Hjørland, Birger, Scerri, E., & Dupré, J. (2011). Forum: The philosophy of classification. *Knowledge Organization*, *38*(1), 9–24.

Hodkinson, T. R. (Ed.). (2011). *Climate change, ecology, and systematics* (1st ed). Cambridge, UK ; New York: Cambridge University Press.

Hopkins, G. W., & Freckleton, R. P. (2002). Declines in the numbers of amateur and professional taxonomists: implications for conservation. *Animal Conservation*, *5*(3), 245–249. https://doi.org/10.1017/S1367943002002299

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., … Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, *1*(5), 16048. https://doi.org/10.1038/nmicrobiol.2016.48

Hull, D. L. (1988). *Science as a process: an evolutionary account of the social and conceptual development of science*. Chicago: University of Chicago Press.

Hull, D. L. (2001). The role of theories in biological systematics. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *32*(2), 221–238. https://doi.org/10.1016/S1369-8486(01)00006-1

Hull, M. S. (2012). Documents and bureaucracy. *Annual Review of Anthropology*, *41*(1), 251–267. https://doi.org/10.1146/annurev.anthro.012809.104953

inference, n. (2016, December). *OED Online*. Oxford University Press. Retrieved from http://www.oed.com/view/Entry/95308

Information Technology Program Group. (1985, July). Proposal for the introduction of a strategy for the use of computers in collection management and scientific database work in the BM(NH). Retrieved from http://www.nhm.ac.uk/research-curation/library/archives/catalogue/DServe.exe?dsqServer=placid&dsqIni=dserve.ini&dsqApp=Archive&dsqCmd=Browse2.tcl&dsqItem=DF%20ZOO%2F269&dsqDb=Catalog&dsqKey=RefNo

instrument, n. (2017, March). *OED Online*. Oxford University Press. Retrieved from http://www.oed.com/view/Entry/97158

Integrated Taxonomic Information System. (2016). ITIS. Retrieved September 1, 2015, from itis.gov

International Association for Plant Taxonomy. (2011, July). International code of nomenclature for algae, fungi, and plants. Retrieved January 11, 2017, from http://www.iapt-taxon.org/nomen/main.php

International Barcode of Life. (2015). International Barcode of Life (iBOL). Retrieved from http://ibol.org/

International Commission on Zoological Nomenclature (Ed.). (1999). *International code of zoological nomenclature* (4th ed). London: International Trust for Zoological Nomenclature, c/o Natural History Museum. Retrieved from http://www.iczn.org/iczn/index.jsp

International Commission on Zoological Nomenclature. (2016a). ICZN Code - glossary. Retrieved November 23, 2017, from http://www.nhm.ac.uk/hosted-sites/iczn/code/index.jsp?booksection=glossary&nfv=true

International Commission on Zoological Nomenclature. (2016b). What is the difference between nomenclature and taxonomy? | International Commission on Zoological Nomenclature. Retrieved February 4, 2017, from http://iczn.org/content/what-difference-between-nomenclature-and-taxonomy

International Commission on Zoological Nomenclature. (2017). Are homonyms across codes permitted, for example between plants and animals? | International Commission on Zoological Nomenclature. Retrieved February 26, 2017, from http://iczn.org/content/are-homonyms-across-codes-permitted-example-between-plants-and-animals

International Union for Conservation of Nature. (2017). The IUCN red list of threatened species. Retrieved January 23, 2016, from http://www.iucnredlist.org/

ISKO Italia. (2004). Integrative levels classification. Retrieved from http://www.iskoi.org/ilc/

ITIS. (2017a). Background information — ITIS. Retrieved March 20, 2017, from https://www.itis.gov/info.html

ITIS. (2017b). ITIS standard report page: Scolytus scolytus. Retrieved March 4, 2017, from https://www.itis.gov/servlet/SingleRpt/SingleRpt?search_topic=TSN&search_value=620499#null

Jens-Erik Mai. (2011). The modernity of classification. *Journal of Documentation*, *67*(4), 710–730. https://doi.org/10.1108/00220411111145061

Jetz, W., McPherson, J. M., & Guralnick, R. P. (2012). Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in Ecology & Evolution*, *27*(3), 151–159. https://doi.org/10.1016/j.tree.2011.09.007

Joint Information Systems Committee. (2017). Janet network. Retrieved March 21, 2017, from https://www.jisc.ac.uk/janet

Joint Steering Committee for Development of RDA, American Library Association, Library Association of Australia, Cataloguing Committee, British Library, Canadian Committee on Cataloguing, … Deutsche Nationalbibliothek. (2015). *RDA: Resource Description & Access*.

Jung, J. J. (2008). Taxonomy alignment for interoperability between heterogeneous virtual organizations. *Expert Systems with Applications*, *34*(4), 2721–2731. https://doi.org/10.1016/j.eswa.2007.05.015

Kalamath Bird Observatory: (2017). Kalamath bird observatory: Species checklist. Retrieved January 24, 2017, from http://www.klamathbird.org/science/methods/54-species-checklist

Kate, K. t. (2002). Global genetic resources: Science and the Convention on Biological Diversity. *Science*, *295*(5564), 2371–2372. https://doi.org/10.1126/science.1070725

Kennedy, J. B., Kukla, R., & Paterson, T. (2005). Scientific names are ambiguous as identifiers for biological taxa: Their context and definition are required for accurate data integration. In B. Ludäscher & L. Raschid (Eds.), *Data Integration in the Life Sciences* (pp. 80–95). Springer Berlin Heidelberg. https://doi.org/10.1007/11530084_8

Kiernan, K. S. (1998). Alfred the Great's burnt Boethius. In G. Bornstein & T. Tinkle (Eds.), *The Iconic Page in Manuscript, Print, and Digital Culture* (pp. 7–32). Ann Arbor, MI: University of Michigan Press. https://doi.org/10.3998/mpub.15786

Kirk, P. (2016, August 30). Personal Interview.

Kirschenbaum, M. G. (2012). *Mechanisms: New media and the forensic imagination*. Cambridge, Mass.; London: The MIT Press.

Kuhn, T. S. (1996). *The Structure of Scientific Revolutions* (3rd ed). Chicago, IL: University of Chicago Press.

Kunze, T., Didžiulis, V., & Roskov, Y. (2013, August). *Proto-GSD in the Catalogue of Life a case study on Mollusca & Platyhelminthes*. Presented at the 48th Annual European Marine Biology Symposium, National University of Ireland, Galway. Retrieved from http://134.213.156.20/sites/default/files/i4lifeposter_Galway_%20Kunze.pdf

Kwasnik, B. H. (2010). Semantic warrant: A pivotal concept in our field. *Knowledge Organization*, *37*(2), 106–110.

Landers, J. (2016, October 18). Big data just got bigger as IBM's Watson meets the Encyclopedia of Life. Retrieved November 1, 2016, from http://www.smithsonianmag.com/smithsonian-institution/ibms-watson-meets-encyclopedia-life-under-new-grant-180960772/

Latham, K. F. (2016, September 15). Fwd: CFP: Resembling science: The unruly object across the disciplines (RBS-Mellon conference, Philadelphia, October 2017).

Le Boeuf, P. (2001). FRBR and Further. *Cataloging & Classification Quarterly*, *32*(4), 15–52. https://doi.org/10.1300/J104v32n04_03

Le Boeuf, P. (2005). FRBR: Hype or Cure-All? Introduction. *Cataloging & Classification Quarterly*, *39*(3–4), 1–13. https://doi.org/10.1300/J104v39n03_01

Lei Zeng, M., & Mai Chan, L. (2004). Trends and issues in establishing interoperability among knowledge organization systems. *Journal of the American Society for Information Science and Technology*, *55*(5), 377–395. https://doi.org/10.1002/asi.10387

Levy, D. M. (2011). *Scrolling Forward: Making Sense of Documents in the Digital Age* (1 edition). Arcade Publishing.

Library of Congress. (2007). Guidelines for Coding Electronic Resources in Leader/06. Retrieved January 14, 2017, from http://www.loc.gov/marc/ldr06guide.html

Locey, K. J., & Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, 201521291. https://doi.org/10.1073/pnas.1521291113

Løvtrup, S. (1987). Phylogenesis, ontogenesis and evolution. *Bolletino Di Zoologia*, *54*(3), 199–208. https://doi.org/10.1080/11250008709355584

Lubetzky, S. (1969). Principles of cataloging. Final report. Phase I: Descriptive cataloging. Retrieved from https://eric.ed.gov/?id=ED031273

Maddison, D. R., Guralnick, R., Hill, A., Reysenbach, A.-L., & McDade, L. A. (2012). Ramping up biodiversity discovery via online quantum contributions. *Trends in Ecology & Evolution*, *27*(2), 72–77. https://doi.org/10.1016/j.tree.2011.10.010

Mai, J. (1999). A postmodern theory of knowledge organization. *Proceedings of the ASIS Annual Meeting*, *36*, 547–56.

Mai, J. (2011). The modernity of classification. *Journal of Documentation*, *67*(4), 710–730. https://doi.org/10.1108/00220411111145061

Mai, J. (2015). About | Global and Local Knowledge Organization. Retrieved October 1, 2015, from http://www.glocalko.info/about/

Marco, F. J. G., & Navarro, M. A. E. (1993). On some contributions of the cognitive sciences and epistemology to a theory of classification. *Knowledge Organization*, *20*(3), 126–132.

Matthias, R. L. F. (2013, July 3). What is a proto-GSD? Retrieved February 7, 2017, from http://blog.catalogueoflife.org/2013/07/why-global-species-databases-and-what.html

Mauro, F., & Hardison, P. D. (2000). Traditional knowledge of indigenous and local communities: International debate and policy initiatives. *Ecological Applications*, *10*(5), 1263–1269. https://doi.org/10.1890/1051-0761(2000)010[1263:TKOIAL]2.0.CO;2

McGann, J. J. (2001). *Radiant textuality: literature after the World Wide Web*. New York: Palgrave.

McGill, B. J., Dornelas, M., Gotelli, N. J., & Magurran, A. E. (2015). Fifteen forms of biodiversity trend in the Anthropocene. *Trends in Ecology & Evolution*, *30*(2), 104–113. https://doi.org/10.1016/j.tree.2014.11.006

Mesibov, B. (2010, May 24). Re: [Taxacom] GBIF: perpetuating probably defunct unpublished names. Retrieved December 2, 2016, from http://taxacom.markmail.org/message/nnqgl4j2qnqxmpqf?q=perpetuating+probably+defunct+unpublished+order:date-forward&page=1#query:perpetuating%20probably%20defunct%20unpublished%20order%3Adate-forward+page:1+mid:ed3ewul7wd2xid6v+state:results

Mesibov, R. (2013). A specialist's audit of aggregated occurrence records. *ZooKeys*, *293*, 1–18. https://doi.org/10.3897/zookeys.293.5111

Mishler, B. D. (2000). Deep phylogenetic relationships among “plants” and their implications for classification. *Taxon*, *49*(4), 661–683. https://doi.org/10.2307/1223970

Montoya, R. D., & Erickson, S. R. (forthcoming). Anachronism in global information systems: the cases of Catalogue of Life and Unicode. *iConference 2017 Proceedings*. https://doi.org/10.9776/16133

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B., & Worm, B. (2011). How Many Species Are There on Earth and in the Ocean? *PLOS Biology*, *9*(8), e1001127. https://doi.org/10.1371/journal.pbio.1001127

Morris, S. A., & Van der Veer Martens, B. (2009). Mapping research specialties. *Annual Review of Information Science and Technology*, *42*(1), 213–295. https://doi.org/10.1002/aris.2008.1440420113

Müller-Wille, S., & Charmantier, I. (2012). Natural history and information overload: The case of Linnaeus. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 4–15. https://doi.org/10.1016/j.shpsc.2011.10.021

New Zealand Organisms Register. (2016). Retrieved February 10, 2017, from http://www.nzor.org.nz/

OED. (2017). contingent, adj. and n. *OED Online*. Oxford University Press. Retrieved from http://www.oed.com/view/Entry/40248

Office of the Secretary, Catherine Hawcker. (1998, April 16). 4/16/98: Integrated Taxonomic Information System Partnership to get Hammer Award. Retrieved February 27, 2017, from http://govinfo.library.unt.edu/npr/library/news/041698.html

Olson, H. (2002). *The power to name: Locating the limits of subject representation in libraries*. Dordrecht; London: Springer.

Orrell, Thomas M., & Schalk, P. (2016, April). *The Catalogue of Life*. Presentation presented at the Catalogue of Life Annual Symposium, Crete, Greece.

Otlet, P., & Rayward, W. B. (1990). *International organisation and dissemination of knowledge: selected essays of Paul Otlet*. Amsterdam ; New York: Elsevier.

P. Bryan Heidorn. (2008). Shedding light on the dark data in the long tail of science. *Library Trends*, *57*(2), 280–299. https://doi.org/10.1353/lib.0.0036

Page, R. D. M. (2012). Space, time, form: viewing the Tree of Life. *Trends in Ecology & Evolution*, *27*(2), 113–120. https://doi.org/10.1016/j.tree.2011.12.002

Page, R., & Michener, B. (Eds.). (2012). *Ecological and evolutionary informatics* (Vol. 27: 2). Elsevier B.V.

Page [@rdmpage], R. (2016a, August 17). .@taxonbytes @timrobertson100 @aeolid Another issue is that @GBIF is an aggregation of sources that maybe themselves be aggregations... [microblog]. Retrieved March 4, 2017, from https://twitter.com/rdmpage/status/765812591683837952?ref_src=twsrc%5Etfw

Page [@rdmpage], R. (2016b, August 17). @taxonbytes @timrobertson100 @aeolid Personally I'd like @GBIF to take more "ownership" of data quality, but that's politically tricky [Microblog]. Retrieved March 4, 2017, from https://twitter.com/rdmpage/status/765813284297736197?ref_src=twsrc%5Etfw

Page [@rdmpage], R. (2016c, August 17). .@taxonbytes @timrobertson100 @GBIF @aeolid And some data we do have is poor (e.g., @catalogueoflife has mangled butterfly names)

[microblog]. Retrieved March 4, 2017, from https://twitter.com/rdmpage/status/765810885390729220?ref_src=twsrc%5Etfw

Page [@rdmpage], R. (2016d, August 17). .@taxonbytes @timrobertson100 @GBIF @aeolid ...so fixing "at source" becomes problematic [microblog]. Retrieved March 4, 2017, from https://twitter.com/rdmpage/status/765812929614741504?ref_src=twsrc%5Etfw

Pape [@fleshflies], T. (2016, November 10). "Names in November" -- in Leiden with people from Species2000, CoL and @GBIF [Social Media].

Pape, T., & Thompson, F. C. (2017). Systema Dipterorum. Retrieved April 1, 2017, from http://www.diptera.org/

Parr, C. S., Guralnick, R., Cellinese, N., & Page, R. D. M. (2012). Evolutionary informatics: unifying knowledge about the diversity of life. *Trends in Ecology & Evolution*, *27*(2), 94–103. https://doi.org/10.1016/j.tree.2011.11.001

Parr, C. S., Lee, B., Campbell, D., & Bederson, B. B. (2004). Visualizations for taxonomic and phylogenetic trees. *Bioinformatics*, *20*(17), 2997–3004. https://doi.org/10.1093/bioinformatics/bth345

Paton, A. (2016, August 24). Personal Interview.

Patterson, D. (2009, July 18). Re: [Taxacom] Catalogue of Life (CoL) management classification draft document. Retrieved February 21, 2017, from http://taxacom.markmail.org/search/?q=document#query:document+page:2+mid:crh25yoge4viij7h+state:results

Patterson, D. J., Cooper, J., Kirk, P. M., Pyle, R. L., & Remsen, D. P. (2010). Names are key to the big new biology. *Trends in Ecology & Evolution*, *25*(12), 686–691. https://doi.org/10.1016/j.tree.2010.09.004

Patterson, D. J., Egloff, W., Agosti, D., Eades, D., Franz, N., Hagedorn, G., … Remsen, D. P. (2014). Scientific names of organisms: attribution, rights, and licensing. *BMC Research Notes*, *7*, 79. https://doi.org/10.1186/1756-0500-7-79

Patterson, D. J., Remsen, D., Marino, W. A., Norton, C., & Page, R. (2006). Taxonomic indexing—Extending the role of taxonomy. *Systematic Biology*, *55*(3), 367–373. https://doi.org/10.1080/10635150500541680

Patterson, D., Mozzherin, D., Shorthouse, D., & Thessen, A. (2016). Challenges with using names to link digital biodiversity information. *Biodiversity Data Journal*, *4*, e8080. https://doi.org/10.3897/BDJ.4.e8080

Pensoft. (2017). Pensoft--About. Retrieved March 4, 2017, from http://pensoft.net/about

Petrak, F. (1969). *Index of fungi, 1920-39: List of new species and varieties of fungi, new combinations and names published* (Vols. 1–10). Kew: Commonwealth Mycological Institute.

Phytokeys. (2017). PhytoKeys. Retrieved March 4, 2017, from http://phytokeys.pensoft.net/articles.php?id=

Pietsch, T. (2015, Winter). *Biodiversity: Methods and goals of systematics, phenetics, and cladistics*. Presented at the Biology of Fishes (Course): Fish/Biol 311, University of Washington. Retrieved from http://courses.washington.edu/fish311/FISH%20311%20files/06-Systematics.pdf

Pilsk, S. C., Kalfatovic, M. R., & Richard, J. M. (2016). Unlocking Index Animalium: From paper slips to bytes and bits. *ZooKeys*, *550*, 153–171. https://doi.org/10.3897/zookeys.550.9673

Plassard, M.-F. (Ed.). (1998). *Functional Requirements for Bibliographic Records, Final Report* (Vol. 19). Munchen, Germany: International Federation of Library Associations.

Plazi. (2017). About Plazi. Retrieved March 4, 2017, from http://plazi.org/about/about-plazi/

Plutarch. (2009). Theseus. Retrieved January 28, 2017, from http://classics.mit.edu/Plutarch/theseus.html

Podani, J. (2013). Tree thinking, time and topology: comments on the interpretation of tree diagrams in evolutionary/phylogenetic systematics. *Cladistics*, *29*(3), 315–327. https://doi.org/10.1111/j.1096-0031.2012.00423.x

Pyle, R. (2016). Towards a Global Names Architecture: The future of indexing scientific names. *ZooKeys*, *550*, 261–281. https://doi.org/10.3897/zookeys.550.10009

Pyle, R. L. (2008, October). *Names, concepts, codes and lots of confusion: An introduction to names of taxonomic organisms*. Presented at the Biodiversity Information Standards (TDWG) Annual Conference 2008, Fremantle (Perth), Australia. Retrieved from http://www.tdwg.org/fileadmin/2008conference/slides/Pyle_03_01_Taxonomic_Names_of_Organisms.swf

Queiroz, K. de, & Gauthier, J. (1992). Phylogenetic taxonomy. *Annual Review of Ecology and Systematics*, *23*, 449–480.

Queiroz, K. D. (1998). Endless forms: Species and speciation. In *The General Lineage Concept of Species, Species Chteria, and the Process of Speciation A Conceptual Unification and Terminological Recommendations* (pp. 57–75). Oxford, England: Oxford University Press.

Ranganathan, S. R., & Gopinath, M. A. (2006). *Prolegomena to library classification* (3rd edition). Bangalore: Ess Ess Publications.

Rapini, A. (2014). Introduction to Botanical Taxonomy. In U. P. Albuquerque, L. V. F. Cruz da Cunha, R. F. P. de Lucena, R. R. N. Alves, L. V. F. C. da Cunha, & R. F. P. de Lucena (Eds.), *Methods and Techniques in Ethnobiology and Ethnoecology* (pp. 123–139). Springer New York. https://doi.org/10.1007/978-1-4614-8636-7_9

Rayward, W. B. (1997). The origins of information science and the International Institute of Bibliography/International Federation for Information and Documentation (FID). *Journal of the American Society for Information Science*, *48*(4), 289–300. https://doi.org/10.1002/(SICI)1097-4571(199704)48:4<289::AID-ASI2>3.0.CO;2-S

Rayward, W. B., La Fontaine, H., & Otlet, P. (2010). *Mundaneum: Archives of knowledge*. Urbana-Campaign, Ill.: Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. Retrieved from https://www.ideals.illinois.edu/handle/2142/15431

Rees, T. (2009, July 21). Re: [Taxacom] Catalogue of Life (CoL) management classification draft document. Retrieved December 2, 2016, from http://taxacom.markmail.org/search/?q=document#query:document+page:2+mid:vkb6h7kzh5qydnmq+state:results

Reichhardt, T. (1999). Catalogue of Life could become reality. *Nature*, *399*(6736), 519–519. https://doi.org/10.1038/21051

Remsen, D. (2010, May 2). Re: [Taxacom] GBIF: perpetuating probably defunct unpublished names. Retrieved December 2, 2016, from http://taxacom.markmail.org/message/rx4bmoywcsb4g653?q=perpetuating+probably+defunct+unpublished

Remsen, D. (2016). The use and limits of scientific names in biological informatics. *ZooKeys*, *550*, 207–223. https://doi.org/10.3897/zookeys.550.9546

Remsen, D. P. (2010). The all genera index strategies for managing the big index of all scientific names. In A. Polaszek (Ed.), *Systema Naturae 250 - The Linnaean Ark*. CRC Press. Retrieved from http://www.crcnetbase.com/doi/book/10.1201/EBK1420095012

Renear, A. H., Sacchi, S., & Wickett, K. M. (2010). Definitions of dataset in the scientific and technical literature. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–4. https://doi.org/10.1002/meet.14504701240

Rheinberger, H.-J. (2010). *On historicizing epistemology: an essay*. Stanford, Calif: Stanford University Press.

Ribes, D., & Finholt, T. (2009). The long now of technology infrastructure: Articulating tensions in development. *Journal of the Association for Information Systems*, *10*(5). Retrieved from http://aisel.aisnet.org/jais/vol10/iss5/5

Rick Szostak. (2008). Classification, interdisciplinarity, and the study of science. *Journal of Documentation*, *64*(3), 319–332. https://doi.org/10.1108/00220410810867551

Robertson, T. (2016, September 16). Personal Interview.

Roskov, Y. (2016a, January 20). Personal Interview.

Roskov, Y. (2016b, February 22). Personal Interview.

Roskov, Y. (2016c, March 31). Personal Interview.

Roskov, Y. (2017, February 7). Personal Interview.

Royal Botanic Gardens, Kew. (2017a). Asystasia nemorum Nees (HerbWeb, Details Page). Retrieved March 3, 2017, from http://specimens.kew.org/herbarium/K000885513

Royal Botanic Gardens, Kew. (2017b). Index Fungorum - Search Page. Retrieved January 21, 2017, from http://www.indexfungorum.org/names/names.asp

Royal Botanic Gardens, Kew. (2017c). Index Fungorum--Home Page. Retrieved December 20, 2016, from http://www.indexfungorum.org/

Royal Botanic Gardens, Kew. (2017d). Kew's science strategy | Science & conservation At Kew. Retrieved February 28, 2017, from http://www.kew.org/kew-science/kews-science-strategy

Royal Botanic Gardens, Kew. (2017e). The International Plant Names Index - home page. Retrieved January 21, 2017, from http://www.ipni.org/index.html

Ruggiero, M. A., Gordon, D. P., Orrell, T. M., Bailly, N., Bourgoin, T., Brusca, R. C., … Kirk, P. M. (2015a). A higher level classification of all living organisms. *PLOS ONE*, *10*(4), e0119248. https://doi.org/10.1371/journal.pone.0119248

Ruggiero, M. A., Gordon, D. P., Orrell, T. M., Bailly, N., Bourgoin, T., Brusca, R. C., … Kirk, P. M. (2015b). S1 Appendix. List of sources consulted for proposed higher level classification of all living organisms. *PLOS ONE*, *10*(4), e0119248. https://doi.org/10.1371/journal.pone.0119248

SAADA. (2016, January 20). Digitization Day @ The Pio Pico Branch Library [Text]. Retrieved January 12, 2017, from https://www.saada.org/losangeles

SæTher, O. A. (1979). Underlying synapomorphies and anagenetic analysis. *Zoologica Scripta*, *8*(1–4), 305–312. https://doi.org/10.1111/j.1463-6409.1979.tb00644.x

Sahagun, B. de. (2012). *Florentine Codex: Book 11: Book 11: Earthly Things*. (C. E. Dibble & A. J. O. Anderson, Trans.) (2 Blg Rev edition). University of Utah Press.

Schalk, P. (2016a, September 7). Personal Interview.

Schalk, P. (2016b, September 8). Personal Interview.

Scheffers, B. R., Joppa, L. N., Pimm, S. L., & Laurance, W. F. (2012). What we know and don't know about Earth's missing biodiversity. *Trends in Ecology & Evolution*, *27*(9), 501–510. https://doi.org/10.1016/j.tree.2012.05.008

schema, n. (2016, December). *OED Online*. Oxford University Press. Retrieved from http://www.oed.com/view/Entry/172307

Schwartz, J. M., & Cook, T. (2002). Archives, records, and power: The making of modern memory. *Archival Science*, *2*(1–2), 1–19. https://doi.org/10.1007/BF02435628

Science of Collaboratories (Home). (2010, 2011). Retrieved from http://soc.ics.uci.edu/

Scott, B., & Smith, V. (2017). Welcome - Data portal. Retrieved March 21, 2017, from http://data.nhm.ac.uk/

Seberg, O., Droege, G., Barker, K., Coddington, J. A., Funk, V., Gostel, M., … Smith, P. P. (2016). Global Genome Biodiversity Network: saving a blueprint of the Tree of Life – a botanical perspective. *Annals of Botany*, *118*(3), 393–399. https://doi.org/10.1093/aob/mcw121

Seddon, J., & Srinivasan, R. (2014). Information and ontologies: Challenges in scaling knowledge for development. *Journal of the Association for Information Science and Technology*, *65*(6), 1124–1133. https://doi.org/10.1002/asi.23000

Short, T. L. (2007). *Peirce's theory of signs*. Cambridge ; New York: Cambridge University Press.

Skouvig, L. (2015, August). *Information culture: Shaping information*. Conference Paper presented at the Global and Local Knowledge Organization, Copenhagen, Denmark. Retrieved from http://www.glocalko.info/

Slota, S., & Bowker, G. C. (2015). On the value of "useless data": Infrastructures, biodiversity, and policy. Retrieved from https://www.ideals.illinois.edu/handle/2142/73663

Smiraglia, R. P. (2002). Further reflections on the nature of "a work": an introduction. *Cataloging & Classification Quarterly*, *33*(3–4), 1–11. https://doi.org/10.1300/J104v33n03_01

Smiraglia, R. P. (2003). The history of "the work" in the modern catalog. *Cataloging & Classification Quarterly*, *35*(3–4), 553–567. https://doi.org/10.1300/J104v35n03_13

Smiraglia, R. P. (2007). Two kinds of power: insight into the legacy of Patrick Wilson. In *Information Sharing in a Fragmented World: Crossing Boundaries*. McGill University, Montreal, Quebec: Canadian Association for Information Science.

Smiraglia, R. P. (2012). Be careful what you wish for: FRBR, some lacunae, a review. *Cataloging & Classification Quarterly*, *50*(5–7), 360–368. https://doi.org/10.1080/01639374.2012.682254

Smith, V. (2016, August 17). Personal Interview.

Smithsonian Institution. (2017). Global Genome Initiative. Retrieved January 20, 2017, from https://ggi.si.edu/

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy; the principles and practice of numerical classification*. San Francisco: W. H. Freeman.

Society of American Archivists. (2017a). Archives. Retrieved April 3, 2017, from http://www2.archivists.org/glossary/terms/a/archives

Society of American Archivists. (2017b). Postcustodial theory of archives. Retrieved January 12, 2017, from http://www2.archivists.org/glossary/terms/p/postcustodial-theory-of-archives

Sokal, R. R. (1966). Numerical Taxonomy. *Scientific American*, *215*(6), 106–116.

Species 2000. (2014, September 23). Catalogue of Life -- Standard dataset (Version 7). Species 2000. Retrieved from http://www.catalogueoflife.org/sites/default/files/datafiles/2014_CoL_Standard_Dataset_v7_23Sep2014.pdf

Species 2000. (2015a). Catalogue of Life. Retrieved July 20, 2015, from http://www.catalogueoflife.org/

Species 2000. (2015b). Catalogue of Life -- About. Retrieved September 1, 2015, from http://www.catalogueoflife.org/content/about

Species 2000. (2015c). Catalogue of Life -- Browse tree. Retrieved September 1, 2015, from http://www.catalogueoflife.org/col/browse/tree

Species 2000. (2015d). Catalogue of Life -- Source databases. Retrieved September 1, 2015, from http://www.catalogueoflife.org/col/info/databases

Species 2000. (2015e). Contributors. Retrieved February 10, 2017, from http://catalogueoflife.org/content/contributors

Species 2000. (2015f). User guide | Catalogue of Life. Retrieved January 14, 2017, from http://www.catalogueoflife.org/content/user-guide

Species 2000. (2016a). 2012 annual checklist: Source databases. Retrieved January 6, 2017, from http://www.catalogueoflife.org/annual-checklist/2012/info/databases

Species 2000. (2016b). 2013 annual checklist: Source databases. Retrieved January 6, 2017, from http://www.catalogueoflife.org/annual-checklist/2013/info/databases

Species 2000. (2016c). 2016 annual checklist : Classification, estimates & extinct taxa. Retrieved February 7, 2017, from http://www.catalogueoflife.org/annual-checklist/2016/info/hierarchy

Species 2000. (2016d). 2016 annual checklist : Taxonomic tree. Retrieved February 8, 2017, from http://www.catalogueoflife.org/annual-checklist/2016/browse/tree/id/f2e91eb663bb9950c972474f65d2693b

Species 2000. (2016e). Catalogue of Life - 2016 Annual Checklist : The 2016 Annual Checklist. Retrieved January 6, 2017, from http://www.catalogueoflife.org/annual-checklist/2016/info/ac

Species 2000. (2016f). Contributing your data | Catalogue of Life. Retrieved January 8, 2017, from http://catalogueoflife.org/content/contributing-your-data

Species 2000. (2016g). What's New? In the Catalogue of Life, 23rd December 2016. Retrieved January 14, 2017, from http://www.catalogueoflife.org/col/info/special

Species 2000. (2017a). Amphinome alba Baird in McIntosh, 1895 (species entry). Retrieved January 29, 2017, from http://www.catalogueoflife.org/col/details/species/id/b61efbba6ea5196deb7b808e25135a5c/source/tree

Species 2000. (2017b). Catalogue of Life - 27th February 2017 : Classification, estimates & extinct taxa. Retrieved March 20, 2017, from http://www.catalogueoflife.org/col/info/hierarchy

Species 2000. (2017c). Frequently asked questions | Catalogue of Life. Retrieved March 4, 2017, from http://catalogueoflife.org/content/frequently-asked-questions#5

Species 2000. (2017d). Previous versions of annual checklist. Retrieved January 6, 2017, from http://catalogueoflife.org/content/previous-versions-annual-checklist

Species 2000. (2017e). Scolytus scolytus Wood & Bright , 1992 (species entry). Retrieved March 4, 2017, from http://www.catalogueoflife.org/annual-checklist/2016/details/species/id/d4fd956e302d8f5e581d2ae71f426bc5

Species 2000 China Node. (2016). Catalogue of Life, China. Retrieved February 10, 2017, from http://www.sp2000.org.cn/

Species 2000 Secretariat. (2015a). Species 2000--About. Retrieved September 1, 2015, from http://www.sp2000.org/about-0

Species 2000 Secretariat. (2015b). Species 2000--Home. Retrieved September 1, 2015, from sp2000.org

Specimen list, n.d., color-coded list of specimens. (1955, 1969).

Srinivasan, R., Boast, R., Furner, J., & Becvar, K. M. (2009). Digital museums and diverse cultural knowledges: moving past the traditional catalog. *The Information Society*, *25*(4), 265–278. https://doi.org/10.1080/01972240903028714

Srinivasan, R., & Huang, J. (2005). Fluid ontologies for digital museums. *International Journal on Digital Libraries*, *5*(3), 193–204. https://doi.org/10.1007/s00799-004-0105-9

Srinivasan, R., Pepe, A., & Rodriguez, M. A. (2009). A clustering-based semi-automated technique to build cultural ontologies. *Journal of the American Society for Information Science & Technology*, *60*(3), 608–620.

Standing Committee of the IFLA Section on Cataloguing. (2009, February 1). Functional Requirements for Bibliographic Records (FRBR): Final report. International Federation of Library Associations and Institutions.

Svenonius, E. (2004). The epistemological foundations of knowledge representations. Retrieved from https://www.ideals.illinois.edu/handle/2142/1691

Svenonius, E. (2009). *The intellectual foundation of information organization* (First MIT Press paperback ed). Cambridge, Mass.: MIT Press.

Szostak, R. (2015). A pluralistic approach to the philosophy of classification. *Library Trends*, *63*(3), 591–614. https://doi.org/10.1353/lib.2015.0007

Tanselle, G. T. (1980). The concept of "ideal copy." *Studies in Bibliography*, *33*, 18–53.

Taylor, H. A. (1982). The collective memory: archives and libraries as heritage. *Archivaria*, *15*, 118–130.

Teckelmann, R., Reich, C., & Sulistio, A. (2011). Mapping of cloud standards to the taxonomy of interoperability in IaaS. In *2011 IEEE Third International Conference on Cloud Computing Technology and Science (CloudCom)* (pp. 522–526). https://doi.org/10.1109/CloudCom.2011.78

Tennis, J. T. (2002). Subject ontogeny: Subject access through time and the dimensionality of classification. *Challenges in Knowledge Representation and Organization for the 21st Century: Integration of Knowledge across Boundaries: Proceedings of the Seventh International ISKO Conference*, *8*, 54–59.

Tennis, J. T. (2006). Function, purpose, predication, and context of information organization frameworks. *Advances in Knowledge Organization*, *10*, 303–310.

Tennis, J. T. (2008). Epistemology, theory, and methodology in Knowledge Organization: toward a classification, metatheory, and research framework. *Knowledge Organization*, *35*(3/2), 102–112.

Tennis, J. T. (2012). The strange case of eugenics: A subject's ontogeny in a long-lived classification scheme and the question of collocative integrity. *Journal of the American*

Society for Information Science and Technology, *63*(7), 1350–1359. https://doi.org/10.1002/asi.22686

Tennis, J. T. (2014). *Emerging concepts in ontogenic analysis*. Poster presented at the iSchool Research Fair, Seattle, WA. Retrieved from https://digital.lib.washington.edu:443/researchworks/handle/1773/37974

Tennis, J. T. (2015). Foundational, first-order, and second-order classification theory. *Knowledge Organization*, *42*(4), 244–249.

The International Barcode of Life Project. (2010, 2015). Barcode of life: Identifying species with DNA barcoding. Retrieved July 2, 2015, from http://www.barcodeoflife.org/

Thiele, K., & Yeates, D. (2002). Tension arises from duality at the heart of taxonomy. *Nature*, *419*(6905), 337–337. https://doi.org/10.1038/419337a

Thomas, C. (2009). Biodiversity databases spread, prompting unification call. *Science*, *324*(5935), 1632–1633. https://doi.org/10.1126/science.324_1632

Thompson, N. H. (1987, November 20). Annual survey of computing in BM(NH). Retrieved from http://www.nhm.ac.uk/research-curation/library/archives/catalogue/DServe.exe?dsqServer=placid&dsqIni=dserve.ini&dsqApp=Archive&dsqCmd=Browse2.tcl&dsqItem=DF%20ZOO%2F269&dsqDb=Catalog&dsqKey=RefNo

Thompson, N. H. (1988, March 15). Options for top level strategy for computing in the Museum. Retrieved from http://www.nhm.ac.uk/research-curation/library/archives/catalogue/DServe.exe?dsqServer=placid&dsqIni=dserve.ini&dsqApp=Archive&dsqCmd=Browse2.tcl&dsqItem=DF%20ZOO%2F269&dsqDb=Catalog&dsqKey=RefNo

Thompson, N. H. (1989, March 5). ITPG Report, Draft 2 additions. Retrieved from http://www.nhm.ac.uk/research-curation/library/archives/catalogue/DServe.exe?dsqServer=placid&dsqIni=dserve.ini&dsqApp=Archive&dsqCmd=Browse2.tcl&dsqItem=DF%20ZOO%2F269&dsqDb=Catalog&dsqKey=RefNo

Thorpe, D. H. (1985, July). Background to the development of computing policy and strategy in the British Museum (Natural History). Retrieved from http://www.nhm.ac.uk/research-curation/library/archives/catalogue/DServe.exe?dsqServer=placid&dsqIni=dserve.ini&dsqApp=Archive&dsqCmd=Browse2.tcl&dsqItem=DF%20ZOO%2F269&dsqDb=Catalog&dsqKey=RefNo

Thorpe, S. (2009, July 18). Re: [Taxacom] Catalogue of Life (CoL) management classification draft document - Stephen Thorpe - edu.ku.nhm.mailman.taxacom - MarkMail. Retrieved February 15, 2017, from http://taxacom.markmail.org/search/?q=document#query:document+page:2+mid:oewytokyqj6f3xag+state:results

"Tree of life" for 2.3 million species released. (2015, September 19). Retrieved September 24, 2015, from http://phys.org/news/2015-09-tree-life-million-species.html

Uetz, P. (2016, July 14). The Reptile Database: General Information (and "FAQ"). Retrieved February 7, 2017, from http://www.reptile-database.org/db-info/introduction.html

United Nations. (1997, May 23). UN conference on environment and development. Retrieved March 18, 2017, from http://www.un.org/geninfo/bp/enviro.html

United Nations. (2017). Agenda 21: Sustainable development knowledge platform. Retrieved March 18, 2017, from

https://sustainabledevelopment.un.org/index.php?page=view&type=400&nr=23&menu=35

Utteridge, T. (2016, August 24). Personal Interview.

Van der Hoorn, B. (2016, September 7). Personal Interview.

Vanden Berghe, E., Coro, G., Bailly, N., Fiorellato, F., Aldemita, C., Ellenbroek, A., & Pagano, P. (2015). Retrieving taxa names from large biodiversity data collections using a flexible matching workflow. *Ecological Informatics*, *28*, 29–41. https://doi.org/10.1016/j.ecoinf.2015.05.004

Vaux, F., Trewick, S. A., & Morgan-Richards, M. (2016). Lineages, splits and divergence challenge whether the terms anagenesis and cladogenesis are necessary. *Biological Journal of the Linnean Society*, *117*(2), 165–176. https://doi.org/10.1111/bij.12665

Vellend, M., Baeten, L., Myers-Smith, I. H., Elmendorf, S. C., Beausejour, R., Brown, C. D., … Wipf, S. (2013). Global meta-analysis reveals no net change in local-scale plant biodiversity over time. *Proceedings of the National Academy of Sciences*, *110*(48), 19456–19459. https://doi.org/10.1073/pnas.1312779110

Wade, N. (2017, March 22). Shaking up the dinosaur family tree. *The New York Times*. Retrieved from https://www.nytimes.com/2017/03/22/science/dinosaur-family-tree.html

Walker, K. (2010, May 23). Re: [Taxacom] GBIF: perpetuating probably defunct unpublished names. Retrieved December 2, 2016, from http://taxacom.markmail.org/message/nnqgl4j2qnqxmpqf?q=perpetuating+probably+defunct+unpublished+order:date-forward&page=1

Waterton, C., Ellis, R., & Wynne, B. (2013). *Barcoding nature: shifting cultures of taxonomy in an age of biodiversity loss*. Milton Park, Abingdon, Oxon: Routledge.

Watson, M. F., Lyal, C. H. C., & Pendry, C. (Eds.). (2015). *Descriptive taxonomy: The foundation of biodiversity research*. Cambridge: Cambridge University Press.

What is EOL? - Information and pictures of all species known to science. (2017). Retrieved February 12, 2017, from http://eol.org/about

Wheeler, Q. D. (2004). Taxonomic triage and the poverty of phylogeny. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *359*(1444), 571–583. https://doi.org/10.1098/rstb.2003.1452

Wiley, E. O., & Lieberman, Bruce S. (2011). *Phylogenetics: Theory and practice of phylogenetic systematics* (2 edition). Wiley-Blackwell.

Wilson, P. (1968). *Two kinds of power; an essay on bibliographical control*. Berkeley: University of California Press.

Wilson, P. (1973). Situational relevance. *Information Storage and Retrieval*, *9*(8), 457–471. https://doi.org/10.1016/0020-0271(73)90096-X

Wilson, P. (1977). *Public knowledge, private ignorance: toward a library and information policy*. Westport, Conn: Greenwood Press.

Wilson, P. (1983). *Second-hand knowledge: an inquiry into cognitive authority*. Westport, Conn: Greenwood Press.

Winston, J. E. (1999). *Describing species: practical taxonomic procedure for biologists*. New York: Columbia University Press.

Witteveen, J. (2015). Naming and contingency: the type method of biological taxonomy. *Biology & Philosophy*, *30*(4), 569–586. https://doi.org/10.1007/s10539-014-9459-6

Woodburn, M. (2016, August 30). Personal Interview.

World Register of Marine Mammals (WoRMS). (2015). Species: Micromussa amakusensis (Veron, 1990). Retrieved January 21, 2017, from http://www.marinespecies.org/aphia.php?p=taxdetails&id=578144

World Register of Marine Species. (2017a). Sabella discifera Grube, 1874 (species). Retrieved February 5, 2017, from http://www.marinespecies.org/aphia.php?p=taxdetails&id=130964

World Register of Marine Species. (2017b). Toward a World Register of Marine Species. Retrieved February 6, 2017, from http://www.marinespecies.org/about.php

World Register of Marine Species. (2017c). WoRMS - World Register of Marine Species. Retrieved February 28, 2017, from http://www.marinespecies.org/sponsors.php

Wright, A. (2014). *Cataloging the world: Paul Otlet and the birth of the information age*. Oxford ; New York: Oxford University Press.

Yanega, D. (2016, August 3). Personal Interview.

Zimmer, C. (2016, April 11). Scientists unveil new "tree of life." Retrieved August 28, 2016, from http://www.nytimes.com/2016/04/12/science/scientists-unveil-new-tree-of-life.html?action=click&contentCollection=Science&module=RelatedCoverage&region=EndOfArticle&pgtype=article

ZooKeys. (2017). ZooKeys. Retrieved March 4, 2017, from http://zookeys.pensoft.net/articles.php?id=