Syst. Biol. 69(6):1231–1253, 2020

DOI:10.1093/sysbio/syaa026
Advance Access publication April 16, 2020

Repositories for Taxonomic Data: Where We Are and What is Missing

AURÉLIEN MIRALLES[1,2], TEDDY BRUY[1,2], KATHERINE WOLCOTT[2,3], MARK D. SCHERZ[4,5], DOMINIK BEGEROW[6], BANK BESZTERI[7], MICHAEL BONKOWSKI[8], JANINE FELDEN[9,10], BIRGIT GEMEINHOLZER[11], FRANK GLAW[4], FRANK OLIVER GLÖCKNER[10], OLIVER HAWLITSCHEK[4,12], IVAYLO KOSTADINOV[13], TIM W. NATTKEMPER[14], CHRISTIAN PRINTZEN[15], JASMIN RENZ[16], NATALIYA RYBALKA[17], MARC STADLER[18], TANJA WEIBULAT[13], THOMAS WILKE[19], SUSANNE S. RENNER[2,*], AND MIGUEL VENCES[20]

[1]*Departement Origins and Evolution, Institut Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, 57 rue Cuvier, CP50, 75005 Paris, France;* [2]*Systematic Botany and Mycology, University of Munich (LMU), Menzingerstraße 67, 80638 Munich, Germany;* [3]*National Museum of Natural History, Smithsonian Institution, Washington, DC, USA;* [4]*Department of Herpetology, Zoologische Staatssammlung München (ZSM-SNSB), Münchhausenstraße 21, 81247 München, Germany;* [5]*Department of Biology, Universität Konstanz, Universitätstraße 10, 78464 Konstanz, Germany;* [6]*Department of Geobotany, Ruhr-University Bochum, Universitätsstraße 150, 44780 Bochum, Germany;* [7]*Department of Phycology, Faculty of Biology, University of Duisburg-Essen, Universitätsstraße 2, 45141 Essen, Germany;* [8]*Department of Terrestrial Ecology, Center of Excellence in Plant Sciences (CEPLAS), Terrestrial Ecology, Institute of Zoology, University of Cologne, 50674 Köln, Germany;* [9]*MARUM - Center for Marine Environmental Sciences, University of Bremen, Leobenerstraße 8, 28359 Bremen, Germany;* [10]*Alfred Wegener Institute - Helmholtz Center for Polar- and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany;* [11]*Department of Systematic Botany, Justus Liebig University Gießen, Heinrich-Buff Ring 38, 35392 Giessen, Germany;* [12]*Department of Scientific Infrastructure, Centrum für Naturkunde (CeNak), Universität Hamburg, Martin-Luther-King-Platz 3, 20146 Hamburg, Germany;* [13]*GFBio - Gesellschaft für Biologische Daten e.V., c/o Research II, Campus Ring 1, 28759 Bremen, Germany;* [14]*Biodata Mining Group, Center of Biotechnology (CeBiTec), Bielefeld University, PO Box 100131, 33501 Bielefeld, Germany;* [15]*Department of Botany and Molecular Evolution, Senckenberg Research Institute and Natural History Museum Frankfurt, Senckenberganlage 25, 60325 Frankfurt/Main, Germany;* [16]*Zooplankton Research Group, DZMB – Senckenberg am Meer, Martin-Luther-King Platz 3, 20146 Hamburg, Germany;* [17]*Department of Experimental Phycology and Culture Collection of Algae, University Göttingen, Nikolausberger-Weg 18, 37073 Göttingen, Germany;* [18]*Department Microbial Drugs, Helmholtz Centre for Infection Research (HZI), and German Centre for Infection Research (DZIF), Partner Site Hannover-Braunschweig, Inhoffenstrasse 7, 38124 Braunschweig, Germany;* [19]*Department of Animal Ecology and Systematics, Justus Liebig University Gießen, Heinrich-Buff Ring 26, 35392 Giessen, Germany; and* [20]*Department of Evolutionary Biology, Zoological Institute, Technische Universität Braunschweig, Mendelssohnstraße 4, 38106 Braunschweig, Germany*

**Correspondence to be sent to: Systematic Botany and Mycology, University of Munich (LMU), Menzingerstraße 67, 80638 Munich, Germany; E-mail: renner@lmu.de*

Received 13 November 2019; reviews returned 20 February 2020; accepted 24 March 2020
Associate Editor: Matt Friedman

Abstract.—Natural history collections are leading successful large-scale projects of specimen digitization (images, metadata, DNA barcodes), thereby transforming taxonomy into a big data science. Yet, little effort has been directed towards safeguarding and subsequently mobilizing the considerable amount of original data generated during the process of naming 15,000–20,000 species every year. From the perspective of alpha-taxonomists, we provide a review of the properties and diversity of taxonomic data, assess their volume and use, and establish criteria for optimizing data repositories. We surveyed 4113 alpha-taxonomic studies in representative journals for 2002, 2010, and 2018, and found an increasing yet comparatively limited use of molecular data in species diagnosis and description. In 2018, of the 2661 papers published in specialized taxonomic journals, molecular data were widely used in mycology (94%), regularly in vertebrates (53%), but rarely in botany (15%) and entomology (10%). Images play an important role in taxonomic research on all taxa, with photographs used in >80% and drawings in 58% of the surveyed papers. The use of omics (high-throughput) approaches or 3D documentation is still rare. Improved archiving strategies for metabarcoding consensus reads, genome and transcriptome assemblies, and chemical and metabolomic data could help to mobilize the wealth of high-throughput data for alpha-taxonomy. Because long-term—ideally perpetual—data storage is of particular importance for taxonomy, energy footprint reduction via less storage-demanding formats is a priority if their information content suffices for the purpose of taxonomic studies. Whereas taxonomic assignments are quasifacts for most biological disciplines, they remain hypotheses pertaining to evolutionary relatedness of individuals for alpha-taxonomy. For this reason, an improved reuse of taxonomic data, including machine-learning-based species identification and delimitation pipelines, requires a cyberspecimen approach—linking data via unique specimen identifiers, and thereby making them findable, accessible, interoperable, and reusable for taxonomic research. This poses both qualitative challenges to adapt the existing infrastructure of data centers to a specimen-centered concept and quantitative challenges to host and connect an estimated ≤2 million images produced per year by alpha-taxonomic studies, plus many millions of images from digitization campaigns. Of the 30,000–40,000 taxonomists globally, many are thought to be nonprofessionals, and capturing the data for online storage and reuse therefore requires low-complexity submission workflows and cost-free repository use. Expert taxonomists are the main stakeholders able to identify and formalize the needs of the discipline; their expertise is needed to implement the envisioned virtual collections of cyberspecimens. [Big data; cyberspecimen; new species; omics; repositories; specimen identifier; taxonomy; taxonomic data.]

Taxonomy, the science of documenting, naming, classifying, and understanding the diversity of life on Earth (Simpson 1961; Small 1989; Stuessy et al. 2014), is deeply embedded in evolutionary biology. It also is of direct relevance for documenting and understanding biodiversity dynamics in the face of global change. Since the current system of binomial scientific names was introduced by Linnaeus (1753, 1758), taxonomists have

named about 1.8 million species (Roskov et al. 2019), and an unknown but undoubtedly vast number of species remain unnamed (Wheeler 2007; Mora et al. 2011; Fontaine et al. 2012; Costello et al. 2013a,b; Locey and Lennon 2016; Larsen et al. 2017). With an estimated global holding of 3 billion biological specimens in collections (Brooke 2000) and some 15,000–20,000 species descriptions per year (IISE 2011: numbers for 2006 and 2007 are 16,969 and 18,516, respectively; this study: Fig. 1) taxonomy clearly qualifies as big data science by fulfilling the main criteria of volume, variety, and velocity (De Mauro et al. 2016). Still, initiatives to implement cybertaxonomic approaches in taxonomic publishing (Smith et al. 2013; Penev et al. 2018) have not been widely adopted, and, most importantly, the rate of new species naming has failed to increase, despite the rise of ever more efficient computational and DNA sequencing tools available. One reason is that the basic species diagnosis and description procedure has remained unchanged (Fig. 1, original data in Supplementary Appendix S1 available on Dryad at http://dx.doi.org/10.5061/dryad.fj6q573qd).

Naming a new species not only involves gathering images, measurements, and molecular sequences for a few reference specimens but also a comprehensive comparative study to distinguish the new from the already known. Little effort has been directed toward harvesting the massive amount of original data that is being generated in the species naming process, and it is therefore often not safeguarded in repositories. As in other fields of evolutionary biology, nonmolecular archived data are often incomplete or insufficiently standardized, and therefore not available for reuse (Roche et al. 2015). Furthermore, many taxonomic journals lack mechanisms (and funds) for the maintenance of online supplementary documents with original specimen-based data, and specialized *taxonomic data repositories* are largely lacking, as we will show below.

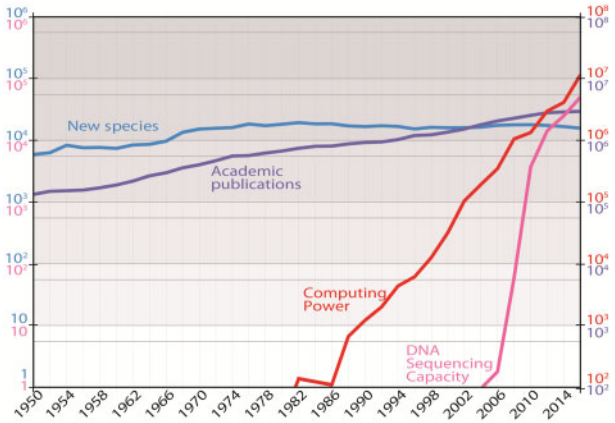

FIGURE 1. Trends over time in taxonomic output (new species named per year) compared to number of academic publications, computing power, and DNA sequencing capacity. Numbers of new species were compiled from the Index to Organism Names (organismnames.com), the International Plant Name Index (ipni.org), and MycoBank (mycobank.org); scientific knowledge is represented as number of academic publications compiled from Scopus (scopus.com); computing power is the number of transistors on silicon chips (Moore's law; data from Rupp 2018); DNA sequencing capacity is the number of Mbp that can be sequenced per 1000 US$. All data presented as 2-year averages (Original data in Supplementary Appendix S1 available on Dryad).

The importance of the availability, connectivity, and management of data in taxonomy is obvious (Gemeinholzer et al. 2020) and is reflected in concepts of cybertaxonomy (Pyle et al. 2008; Winterton 2009; LaSalle et al. 2009; Padial et al. 2010; Balke et al. 2013; Favret 2014; Rosenberg 2014; Stackebrandt and Smith 2019). As claimed by Bik (2017), if we play our cards right, taxonomy could be on the brink of another golden age. Driven by the need to comprehensively explore and document Earth's species (Wheeler et al. 2012a), big advances are being made in building cybertaxonomic infrastructures, especially by digitally mobilizing metadata and images of voucher specimens in biological collections as well as literature, by increasingly registering nomenclatural acts online (Krell 2015), and by building curated databases of species names, diagnoses, and descriptions (Crous et al. 2004; Patterson et al. 2010; Webster 2017). At the moment, for instance, 172 *taxonomic databases* are contributing to the *Catalogue of Life* (Roskov et al. 2019).

Here, we review the data repositories currently available for taxonomic data and describe how improved data management could contribute to improving the inventory of life on Earth. We focus on alpha-taxonomy the purpose of which is to establish an inventory of the past and present species on Earth, combining (i) a fundamental component, grounded in evolutionary biology, which consists of specimen-based species delimitation and (ii) an applied component, which consists of providing a universal communication system to unambiguously communicate about biodiversity. This is achieved via the assignment of a two-part name in Latin ruled by taxon-specific codes of nomenclatures, all of which require (i) designating type material from a collection and (ii) a diagnosis that sets the new taxon apart from the most similar already named taxa. Descriptions are not mandatory in any of the five codes for the simple reason that Linnaeus did not use them, relying instead on concise diagnoses (Renner 2016).

Data for fundamental research in alpha-taxonomy of eukaryotes necessarily are specimen-based. They are therefore not covered in species-based *taxonomic databases* that store information on diagnostic features, synonymy, distribution, phylogeny, traits, or natural history of species. The original analyses carried out for this study show that many established *data repositories* do not meet the requirements of taxonomists for data submission, retrieval, searchability, and reuse.

PROPERTIES AND DIVERSITY OF ALPHA-TAXONOMIC DATA

Historically, taxonomy was based on an essentialist concept, with members of a species assumed to share an essence setting them apart from other species. Today, taxonomy is embedded in evolutionary biology, and species are seen as inferred population-level evolutionary lin-

eages (Mayden 1997; de Queiroz 1998, 2007; Padial et al. 2010). This change of paradigm, however, did not change how other biological disciplines, and most end users of taxonomies, tend to conceive and utilize taxonomic species hypotheses: individual organisms are examined and their traits are considered as representative for the nominal species to which they were assigned by the most recent taxonomist to label or otherwise "identify" the organism in question (Supplementary Appendix S2 available on Dryad). This implies that databases for end users of taxonomy, in science, and society, will be centered on species names: traits, geographic ranges, taxonomy, phylogeny, diagnoses, images, or DNA sequences will primarily be labeled with and retrieved via scientific names and conceptualized as representing the respective species in other research, identification tools, laws, and conservation assessments.

The alpha-taxonomic workflow itself, that is, the elaboration of species hypotheses, follows a different approach. Ideally, multiple individuals are studied to infer "sufficiently" divergent, evolutionarily independent population-level lineages, and based on this evaluation, they are assigned species rank. The species is thus not the basic unit of research, but instead the endpoint and result of a study (Supplementary Appendix S2 available on Dryad). Independent of the species concept and species criteria used, alpha-taxonomic research is centered on individual organisms in order to assess variation and so are the data produced during this research activity.

The unit studied by alpha-taxonomists typically is a *specimen*—either an individual organism, or in the case of paleontology, part thereof, or a cultured isolate composed of multiple, often clonal individuals. Of particular importance are name-bearing *type specimens*, which constitute anchors for assigning a scientific name to a species. Almost universally, these are physical objects (preserved organisms or their parts, metabolically inactive strains, or living, viable cultures) as recommended by all five codes of nomenclature (Amorim et al. 2016; Santos et al. 2016; Renner 2016), although where type specimens are declared lost, images can be used. Fierce disputes revolve around the option of basing new scientific nomina on photographs, videos, or DNA sequences alone (Ceriaco et al. 2016; Thorpe 2017; Krell and Marshall 2017; Garraffoni and Freitas 2017). In mycology, proposals have been put forward to allow DNA sequences alone, even environmental DNA sequences, as a basis for naming new species (Hawksworth et al. 2016) but the majority of mycologists are presently reluctant to accept voucherless species-level taxa to be validly erected (May et al. 2018), also because many comparative DNA sequences available from repositories are insufficiently linked to permanently preserved specimens (Hongsanan et al. 2018; Zamora et al. 2018).

Because physical specimens in collections are not always accessible and deteriorate as they age or as they are destructively sampled for carbon-14 dating, scanning electron microscopy, or DNA isolation, some authors have pushed for the introduction of digital type specimens or *cybertypes* (e.g., Godfray 2007). Such cybertypes would be a complement (not a substitute) to physical types deposited in collections. Representing visual type information online is becoming more widespread (Bosselaers et al. 2010; Wheeler et al. 2012b; Faulwetter et al. 2013; Akkari et al. 2015; Scherz et al. 2016a,b). Wheeler et al. (2012b) suggested that a cybertype should minimally comprise a photo of the holotype and ideally additional photos of the organism in life, as well as detailed photos of important diagnostic characters. The cybertypes of Faulwetter et al. (2013) and Akkari et al. (2015), for example, include microCT scans with iodine, also known as diffusible iodine-based contrast-enhanced computed tomography (diceCT), which were used to create 3D digital models of the external and internal morphology of specimens without permanently damaging them (Gignac et al. 2016). Such *cyberspecimens* (Favret 2014), could be expanded by nonvisual characteristics (e.g. DNA sequences or sound recordings, Fig. 2). Standards for digital representations of specimens are so far lacking, but it is obvious that the cyberspecimen concept, also referred as extended specimens (Cicero et al. 2017; Lendemer et al. 2020), implies digital publication of extensive and diverse data packages connected via unique specimen identifiers.

The data that are generated in taxonomic research—and that would make up a cyberspecimen—are extremely diverse, depending on the organisms studied and the methods used (Table 1). They comprise both metadata and taxonomic data, and in a data management context it is crucial to conceptually distinguish these two categories (Fig. 3). Metadata come in different categories (Riley 2004): in alpha-taxonomy, *specimen metadata* characterize a specimen as a *collection item*, and contextualize it (Leonelli 2014) by providing information on taxonomic assignment (species name, supraspecific ranks), type status, spatial, and temporal origin (collection date and location), and other technical and historical characteristics (collector name, preservation modality or storage coordinates including institution, collection, individual identifier). In contrast, *taxonomic data* are those that characterize the specimen as a *biological entity*. They represent different kinds of raw or encoded data intended to capture or to describe biological characteristics, such as morphological, anatomical, molecular, or behavioral traits. They are most often generated *a posteriori* in the framework of research that includes the specimen, but they can also be generated *in situ* during the collection of the specimen, anticipating future investigations (for instance, pictures taken in the field to document coloration in life).

Data on a specimen comprise both raw data and processed, selected, and encoded data (Fig. 4; Table 2). Specimen metadata can also become important raw material for taxonomic research, for example, when

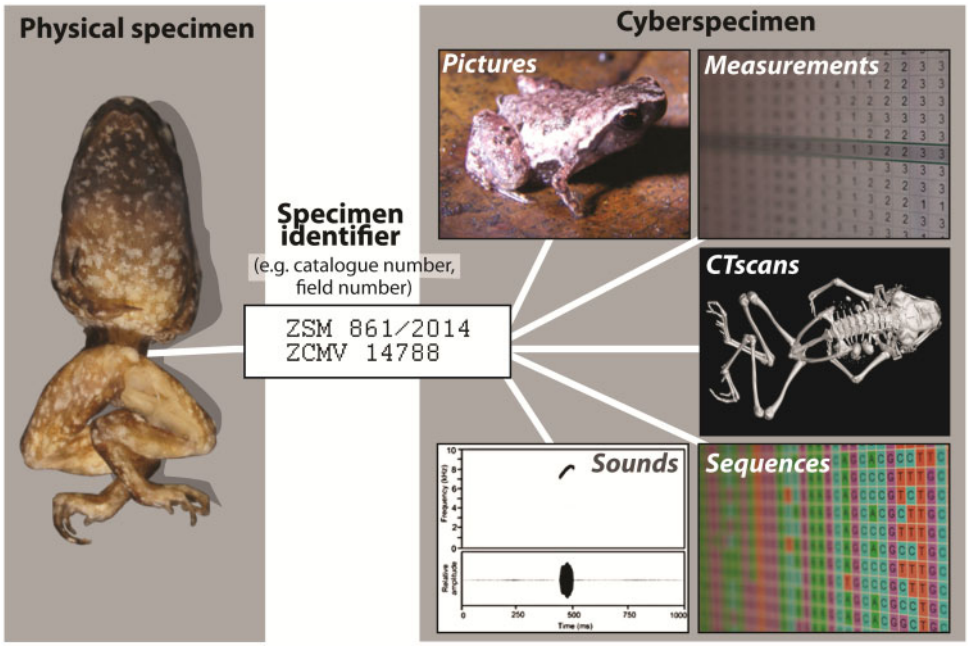

FIGURE 2. Schematic representation of a cyberspecimen: a virtual representation of a physical specimen to which the cyberspecimen is linked by a unique identifier. Primarily, the cyberspecimen consists of a high-resolution digital representation, ideally a 3D image obtained for example, via microCT-scanning or photogrammetry. The cyberspecimen additionally contains all other digital data obtained specifically from the specimen, including photographs of the specimen "in life," morphometric data, genetic (genomic) sequences, sound files, or chemical profiles, all linked by the specimen identifier to the physical/true specimen.

geographical coordinates are used to model and distinguish environmental niches (e.g. Rissler and Apodaca 2007; Cicero et al. 2017), when time of collection is used to characterize phenology, migration behavior, or invasiveness (Chauvel et al. 2006; Miller-Rushing et al. 2006; Grass et al. 2014; Lorieul et al. 2019), or to determine the applicable regulations for access and benefit sharing, which depend on the time when a specimen was acquired by a collection.

The heterogeneous taxonomic data themselves also need to be described with contextual or methodological information. This might include the device, methods, and conditions used for photographic, tomographic, or sound recording (Roch et al. 2016; Köhler et al. 2017), laboratory methods for histological staining or molecular sequencing, or even the sociological context of the data collection (McClellan 2019)—these constitute what one might consider "metadata of taxonomic data" (not the same as specimen metadata). Ideally, these data and metadata must all be accommodated in the archiving process. On the one hand, these intricate requirements suggest that a distributed system of specialized repositories for specific kinds of taxonomic data would be the best approach. On the other hand, it is preferable to adjust the existing infrastructure of established repositories rather than create new ones and to streamline the submission process of diverse data via user-friendly submission portals. The key lies in linking the data to a single specimen for which a *specimen identifier* will be required (Güntsch et al. 2018).

The specimen identifier approach still has to overcome multiple practical problems due to ambiguities in defining what a specimen is (Supplementary Appendix S3 available on Dryad). For instance, in most insect collections, specimens—individual insects in the collection—have no identifying number and usually also lack a catalog that could provide an inventory of specimens. Even type specimens may lack individual specimen identifiers (e.g. Zompro 2005). This is a massive impediment considering an overall estimated 500 million preserved insect specimens in collections (Short et al. 2018). For most of these specimens, the associated metadata are pinned on small labels underneath the specimen and therefore cannot be scanned without labor-intensive unpinning of every specimen. If several specimens have been collected at the same location and time, their metadata will be identical, and distinguishing among these specimens is impossible from the metadata. While it is possible to consider these and other bulk samples as one specimen, problems arise if data (DNA sequences, images, measurements) refer to only one of the individuals included in the bulk, and problems are exacerbated if the bulk is found to contain individuals of different characteristics or even species (see also Nelson et al. 2018).

Many natural history collections are currently digitizing their specimens. For instance, 91% of the 5.5 million plant specimens deposited in the world's largest herbarium (MNHN in Paris) have been photographed at high resolution and made available online in less than a

TABLE 1. Data types used and/or produced in the context of taxonomy, currently or potentially in the future, their predicted storage requirements and main issues to be solved to allow their efficient storage and reuse

	Current use in alpha-taxonomy	Potential and prospective use in taxonomy	Storage requirements (per specimen)	Established specialized repositories	Issues and gaps
Regular images (e.g., .jpeg, .pdf, .tiff)	Regularly used	Images of different kinds will continue to be a main workhorse of taxonomic description and identification; new perspectives by machine-learning character extraction	Moderate to very high, depending on image quality and quantity	Yes (many specialized and generalist repositories will accept images)	Images produced in taxonomic revisions are rarely submitted to repositories; images are often not linked to specimen identifiers
High-resolution images (stacks etc.) (e.g., .tiff)	Increasingly used, e.g. in insects	As with regular images	High to very high	Yes (many specialized and generalist repositories will accept images)	As with regular images
Annotated images	Very rarely used	Documentation of morphology of small-sized organisms (e.g., on a microscopic slide)	High to very high	Only few specialized repositories	Requires development of standards for repositories and submitters
3D microCT, photogrammetry, and laser scanners (e.g., stack of .tiff, polygon mesh such as .ply, .bend, .obj)	Used rarely but regularly, especially in vertebrates. Increasing use in invertebrates.	High importance to visualize internal features of an organism or 3D morphometrics, key method in cyberspecimen approaches	High to very high, depending on storage modality (e.g., polygon mesh vs. raw data) and level of resolution	Yes, several	Requires development of standards for repositories and submitters. See commentary by Hipsley and Sherratt (2019).
DNA sequences (Sanger) (e.g., .fasta, .fastq, .gb)	Regularly used for most organism groups, almost omnipresent in mycological taxonomy	DNA barcodes will continue to drive species identification and discovery, multigene phylogenies important for inferring relationships	Very low to low depending on the number of loci sequenced	Yes, several very well established ones	Sequences deposited in databases are not always curated postsubmission, leading to mismatches after taxonomic changes.
RNAseq (raw) (e.g., .fastq)	Not used	Potentially useful after read mapping and variant calling, but currently rarely used.	High	Yes (e.g., Sequence Read Archive)	No issues

(*Continued*)

TABLE 1. (Continued)

	Current use in alpha-taxonomy	Potential and prospective use in taxonomy	Storage requirements (per specimen)	Established specialized repositories	Issues and gaps
RNAseq (assembly) (e.g., .fasta)	Very rarely used	Valuable source of sequences for phylogenomics and species delimitation	Low	Yes (e.g., Transcriptome Shotgun Assembly Sequence Database)	Assemblies are often not submitted to repositories, although they could be a valuable source of sequences for machine-learning species discovery pipelines
Amplicon (raw) (e.g., .fastq)	Not used	Not straightforward; requires filtering and preprocessing	High	Yes (e.g., Sequence Read Archive)	No issues
Amplicon (consensus OTUs) (e.g., .fasta)	Not used	Metabarcoding data helps ascertaining distribution and ecology of taxa	Low	Not really. Sequences >200 bp could be submitted to GenBank.	OTU consensus sequences from metabarcoding studies are in most cases not submitted to a repository, but could be important for DNA-based assessments of distribution of taxa; targeted and searchable repositories do not exist (GenBank does not accept sequences <200 bp)
Bait capture—raw (e.g., .fastq)	Not used	Only usable after assembly	High	Yes (e.g., Sequence Read Archive)	No issues
Bait capture—assembled (e.g., .fasta)	Rarely used (e.g., sequencing of historical types)	Very valuable source of sequences for phylogenomics and species delimitation—next-generation DNA barcoding	Low to moderate	Yes, well established, same ones as for Sanger sequences	Similar as for Sanger sequences

(*Continued*)

TABLE 1. (Continued)

	Current use in alpha-taxonomy	Potential and prospective use in taxonomy	Storage requirements (per specimen)	Established specialized repositories	Issues and gaps
Genomes—raw (e.g., .fastq)	Not used	Only usable after assembly	Very high	Yes (e.g., Sequence Read Archive)	Similar as for Sanger sequences
Genomes—assembled (e.g., .fasta)	Very rarely used	Valuable source of sequences for phylogenomics and species delimitation	High	Yes	Similar as for Sanger sequences
Maldi-TOF (e.g., .raw, .mzXML, .mzML)	Sometimes used in mycology; commonly in prokaryotes.	Useful for chemotaxonomic approaches	Moderate to very high, depending of storage of spectra vs. raw data.	No	Requires development of standards for repositories and submitters
Near-infrared spectroscopy (e.g., .snirf, .csv, .spc, and many others)	Not used	Possibly useful for "metabolomic barcoding"	Moderate	No	Requires development of standards for repositories and submitters
GC-MS/ (e.g., .raw, .cdf, .D, .mzxml)	Sometimes used in mycology; commonly in prokaryotes.	Useful for chemotaxonomic approaches, e.g., fatty acid profiling in yeasts, and in bacterial taxonomy	Moderate to very high, depending of storage of spectra vs. raw data.	No	Requires development of standards for repositories and submitters; reference databases do exist.
NMR/TLC/HPLC (e.g., .raw , .data, ..cdf)	Rarely used (e.g., TLC and HPLC in lichenology)	Possibly useful for chemotaxonomic approaches	Moderate to very high, depending of storage of spectra vs. raw data.	No	Requires development of standards for repositories and submitters

(*Continued*)

TABLE 1. (Continued)

	Current use in alpha-taxonomy	Potential and prospective use in taxonomy	Storage requirements (per specimen)	Established specialized repositories	Issues and gaps
Sounds (e.g., .wav, .mp3)	Regularly used in sound-producing animals	Very useful for species delimitation of sound-producing animals	Moderate to high, depending on file format and sound duration	Yes	Most repositories do not feature user-friendly submission procedures and often data are not open access
Videos (.avi, .mov, .mp4)	Very rarely used (e.g., to document specific behavior)	Limited value	Moderate to very high, depending on definition and duration	Yes	Extend image databases to accept videos if linked to specimens
Measurements (e.g., .csv, .xls. .txt)	Regularly used	Very useful basic data for diagnosis and identification of species	Very low	No	Requires development of standards for repositories and submitters
2D geometric morphometric data sets (e.g., .csv)	Very rarely used	Increasingly used for resolving species complexes	Very low to low	Yes	No issues
3D geometric morphometric data sets (e.g., .csv)	Very rarely used	Increasingly used for resolving species complexes, especially in combination with microCT scans	Very low to low	Yes	No issues

Note: Note that the second column specifically focuses on the current use of data types in alpha-taxonomic studies (mostly based on our survey reported below), not other taxonomy-related activities such as species identification or phylogenetics. Storage capacity required per specimen: very low (<0.1 MB), low (0.1–1 MB), moderate (1–10 MB), high (10–100 MB), and very high (>100 MB).

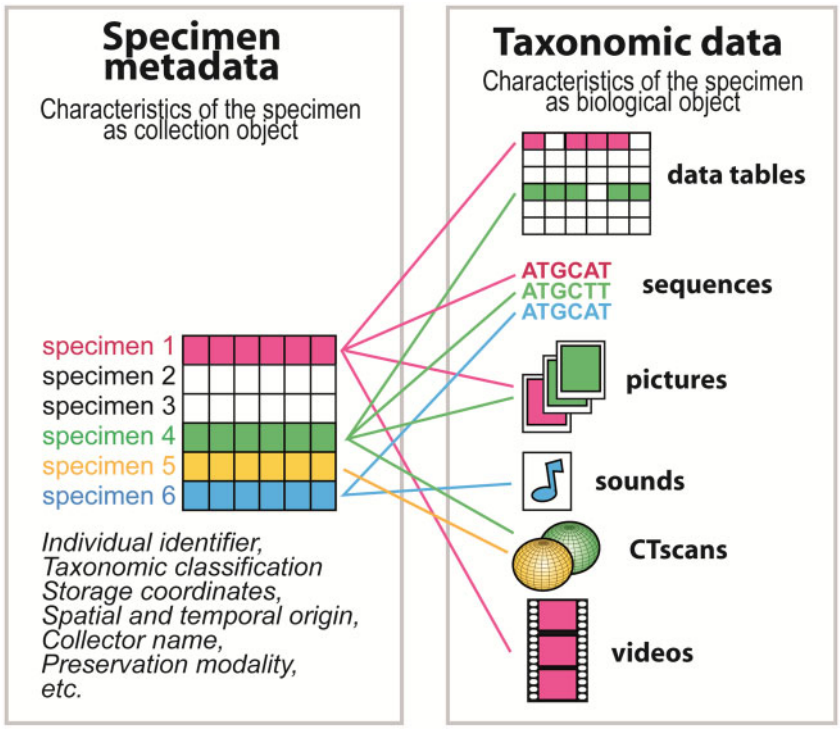

FIGURE 3. Two categories of data linked to a specimen: metadata and taxonomic data. While specimen metadata from museum catalogs are increasingly made digitally available, the scarceness of specialized specimen-based data repositories adapted to the wide range of taxonomic data types is a limitation for the development of digital taxonomy. Additionally, "metadata of taxonomic data" (not shown) are associated with the taxonomic data (e.g., device used, methodology, author name, and date of the measurement).

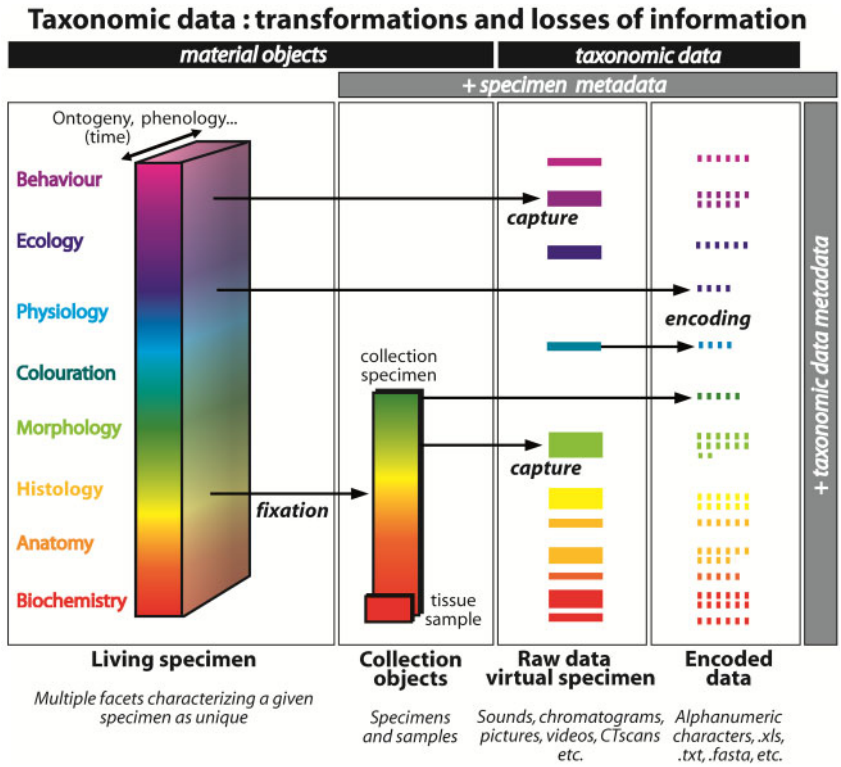

FIGURE 4. Overview of data types, transformations, and specification of information in the process of specimen-based alpha-taxonomic research. Paleontological samples can be considered to be already "fixed" for the purposes of this graphic, by the process of fossilization.

TABLE 2. Properties of different kinds of specimen-based taxonomic data

a. Raw vs. encoded taxonomic data.
These two categories differ by the quantitative and qualitative nature of the information they convey, and consequently by their ease and cost of storage.

Description	**Raw taxonomic data:** One of the multiple facets characterizing a specimen captured by a sensor (e.g., camera, sound recorder, scanner, DNA sequencer). Allows one to represent the different properties of a virtual specimen (e.g., coloration, shape, size, structure, texture, chemical composition, bioacoustics properties).	**Encoded taxonomic data:** Data already interpreted.
Strengths	Free of interpretation. Containing much information. Different forms of encoded data can be extracted, including by future methods that do not yet exist.	Although specific file types may exist, it is always possible (and easy) to translate them into a suite of alphanumeric characters compatible with a universal format (such as .csv). Small files, minimal storage cost.
Weaknesses	Cannot be directly used as input for analyses (need to be interpreted and encoded, either by a human or artificial intelligence). Different and specific storage formats. Large files and high storage cost.	Information restricted to a minimal level. Subjectivity: alternative interpretations (or coding errors) are possible.
Examples	Photographs, sound and video recordings, microCT scans, chromatograms depicting DNA sequences.	Quantitative measurements, qualitative traits encoded numerically or described in natural language, nucleotide or amino acid sequences, morphometric landmarks.

b. Taxonomic data of unique vs. multiple specimens.
These two categories of data differ in the way they are (i) submitted to repositories, (ii) searched for (a particular specimen nested in a multiple-specimen data set has to be detectable using basic search options), and (iii) presented and downloaded on the repository interface. Ideally, for multiple specimen data sets, it should be possible to download either the data measured for a particular specimen only, or the whole data set.

Description	**Unique specimen data:** Data or set of data concerning a single specimen. Most often it consists of raw taxonomic data (see above).	**Multiple-specimens data sets:** Set of data concerning particular trait(s) measured for several specimens. Most often it consists of encoded data.
Strengths	Specific (individual) searches are easy.	Submission of large data set at once.
Weaknesses	Case by case treatment unrealistic where specimen numbers are $>> 10^2$	Stringent search and data extraction might be compromised by an inadequate data archiving process.
Examples	Picture(s) of a specimen, complete mitogenome sequence of a given individual.	Tabular data (.csv), DNA alignment (.fasta, .nex).

decade (Le Bras et al. 2017, constantly updated online at https://edition-humboldt.de), although so far only 16% have field-collecting information (label data) associated with them. Important efforts are also being made on several entomology collections (specimen images and metadata; e.g., Dietrich et al. 2012). So far, however, only an estimated 2% have been digitized (Short et al. 2018). To allow taxonomists to efficiently access, use and reuse these data, individual specimen identifiers are essential (Page 2016; Güntsch et al. 2018), and consequently, priority efforts are usually directed towards providing specimen identifiers to type specimens and accordingly adding labels to the physical types in the collection. Surprisingly, the *International Code of Zoological Nomenclature* (Anonymous 1999) does not require individual identifiers for type specimens.

Quantifying the Kinds of Data Used and Produced in Alpha-taxonomy

To understand which repositories and storage capacities are needed for taxonomic data we quantitatively assessed the number of alpha-taxonomic studies and the kinds of data produced in them. An updated summary of numbers of studies naming new insects, plants, mollusks, fungi, and vertebrates from 1950 to 2016 (Fig. 5) illustrated a noticeable increase after 1966 for insects, with >8000 new species named per year, while in plants, a peak was apparent in the 1980s. Species discovery and naming in fungi has been undergoing a striking increase since 2010 (see also Cannon et al. 2018), whereas for vertebrates numbers have risen more continuously.

Molecular data are at the core of a modern, integrative taxonomy (Padial et al. 2010). To assess their impact, we undertook a systematic search in Web of Science using a combination of search terms to detect alpha-taxonomic studies referring to molecular data during the years 1990–2018 (search terms: molecular, DNA, gene; details in Supplementary Appendix S4 available on Dryad). The results confirm a raise in the explicit use of molecular evidence across all groups (Fig. 6). Mycologists and protistologists mention molecular data in >75% of their taxonomic studies in 2018, whereas this was the case for only 33% of insect and 26% of plant studies. Such an increasing use of DNA sequences in taxonomy likely reflects a growing tendency to take evolutionary concepts into account during the species delimitation process, even if only implicitly.

We next undertook a survey of 4178 alpha-taxonomic studies (published in 2002, 2010, and 2018) that involved scientific naming of species. Each of these was manually screened, and kinds of data used in the respective study were tabulated, along with a series of metadata for each paper. We surveyed the taxonomic journals *Phytotaxa, Zootaxa, Systematic Botany*, and *Mycological Progress*, and six generalist journals with higher-impact factors (*Nature, Science, PNAS, PLoS One, Scientific Reports*, and the *Biological, Botanical* and *Zoological Journal of the Linnean Society*, for alpha-taxonomic studies (Supplementary Appendix S5 available on Dryad). The average publication named 1–2 (fungi, plants, protists, vertebrates) or 3–4 (insects and other invertebrates) new species (Fig. 7; original data in Supplementary Appendix S6 available on Dryad).

In this survey, we more restrictively considered the use of a certain kind of data only if it was explicitly part of the arguments supporting a taxonomic change (usually the description of a new species). The use of molecular evidence, newly generated or from other sources, was similar to our Web of Science survey (Fig. 6). In the specialized taxonomic journals (4113 studies), molecular data were widely used in mycology, but much less so in botany and zoology (Supplementary Appendix S7 available on Dryad): in 2018 papers, DNA sequence analysis was used in 94% of taxonomic studies of fungi, 53% of vertebrates, 15% of plants, and 10% and 14% of insects and other invertebrates (Fig. 7). Surprisingly, even in works on protists, which are difficult to identify morphologically, genetic evidence was used in only 29% of the 66 surveyed papers, although our Web of Science survey suggested otherwise (Fig. 6). Even the frequent DNA use in mycology suggested by our survey may be an overestimate because many fungi are described in other specialized journals not surveyed here, mostly without molecular data. Comparing papers from 2002, 2010, and 2018, an increase in the use of molecular evidence is apparent for all organismal groups (Fig. 7).

Photographic images were used in >80% of the papers in all categories in 2018, whereas other sets of data (extensive morphometric data sets or 3D-imagery) were only rarely used and almost restricted to studies on vertebrates. Specifically, in the entire set of 4113 papers, only 17 studies used microCT-scanning, 2 used synchrotron-based visualization, 6 used other kinds of 3D-visualization, 14 used X-ray images, and 1 used videos. Besides macroscopic photos, microscopy and microscopy-produced images were used frequently: 670 (16%) studies used electron microscopy (SEM or TEM) and 709 (17%) used light microscopy. Classical drawings were part of 2371 (58%) of the 4113 studies.

Genome-scale data sets (e.g., RADseq, Sequence capture, RNAseq, full genomes) in 2018 were only used in one paper in mycology (a draft genome), and not at all in zoology or in botany. Similarly, metabolomics data were rare in the surveyed papers in 2018: one publication using NIR spectra in entomology, one using NMR spectra in mycology, and one using peptide fingerprints in vertebrate zoology.

Several other kinds of molecular data were used in a moderate proportion of the 4113 papers: cytological techniques from cell descriptions to flow-cytometric determination of ploidy and genome size ($n=329$), karyotypes ($n=34$), fragment analysis (microsatellites, AFLP, RFLP, $n=10$), allozymes ($n=3$), and chemotaxonomic approaches including analysis of cuticular hormones or metabolites ($n=17$) and GC-MS or HPLC metabolite profiles ($n=4$).

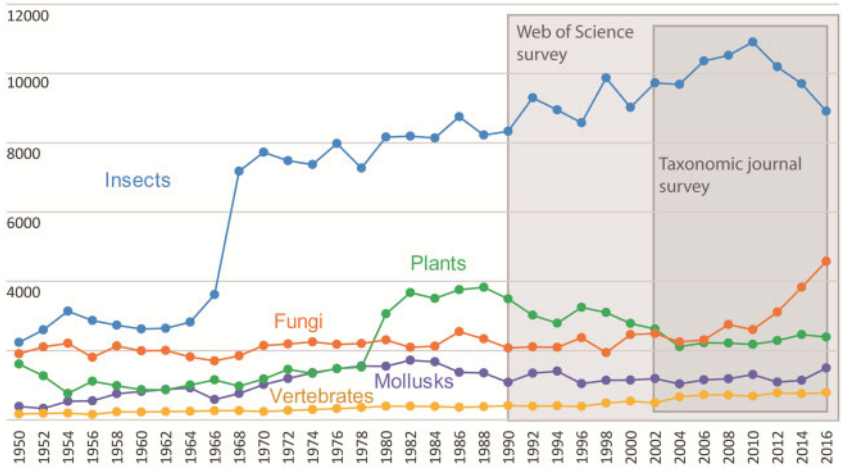

FIGURE 5. Species named per year for the study period. Insects and mollusks (ION—organismnames.com), fungi (MycoBank—mycobank.org), plants (IPNI—ipni.org), vertebrates (compiled from Eschmeyer's Catalog of Fishes, Amphibian Species of the World: Frost 2019, Reptile Database, Howard, and Moore Bird Checklist: Christidis 2018, Mammal Diversity Database; all accessed in March 2019: calacademy.org/scientists/projects/eschmeyers-catalog-of-fishes, research.amnh.org/vz/herpetology/amphibia/, reptile-database.org, mammaldiversity.org). The gray-shaded windows indicate the time frames for which our surveys of data types were carried out. Vertebrate numbers refer to currently accepted species, whereas for the other taxa, also species currently considered as synonyms are included. Furthermore, the ION data (insects) also include subspecies. Note that paleontological studies were excluded from our survey.

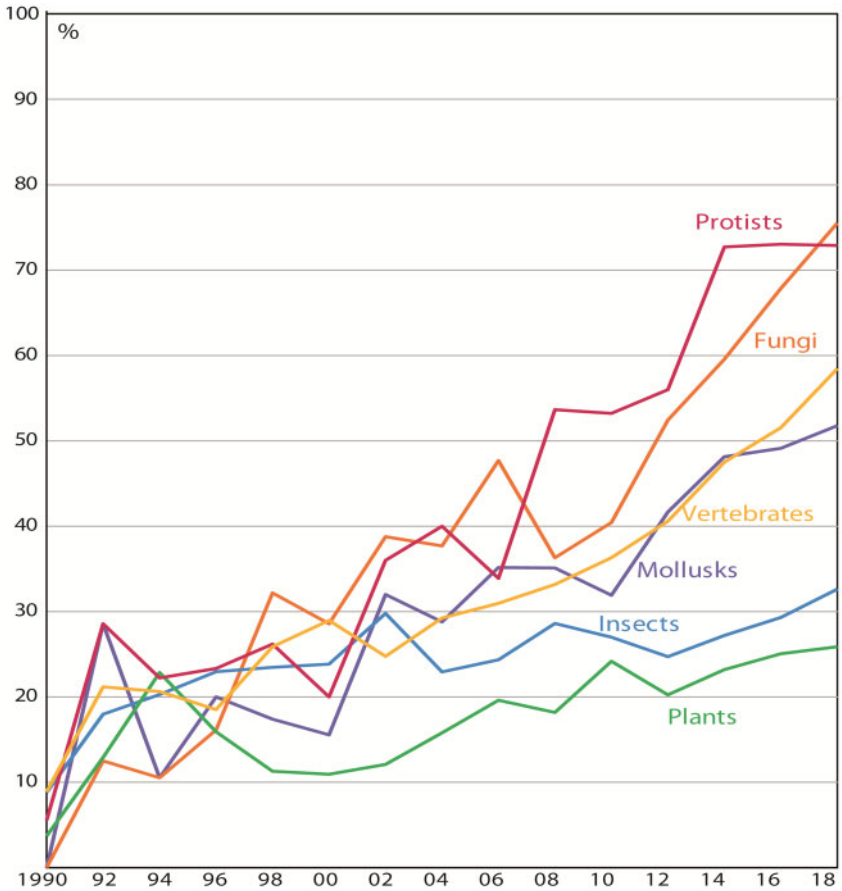

FIGURE 6. Comparison of the frequency of use of molecular data in taxonomic studies naming new species, in various groups of organisms, given as proportion of taxonomic papers retrieved from a semantic search on Web of Science, measured every 2 years. Molecular data were considered as contributing to every study based on a combination of search terms (cf. details in the Supplementary Appendix S4 available on Dryad). Data do not necessarily reflect absolute numbers due to inaccuracy involved with keyword searches but primarily serve as a comparison among organism groups.

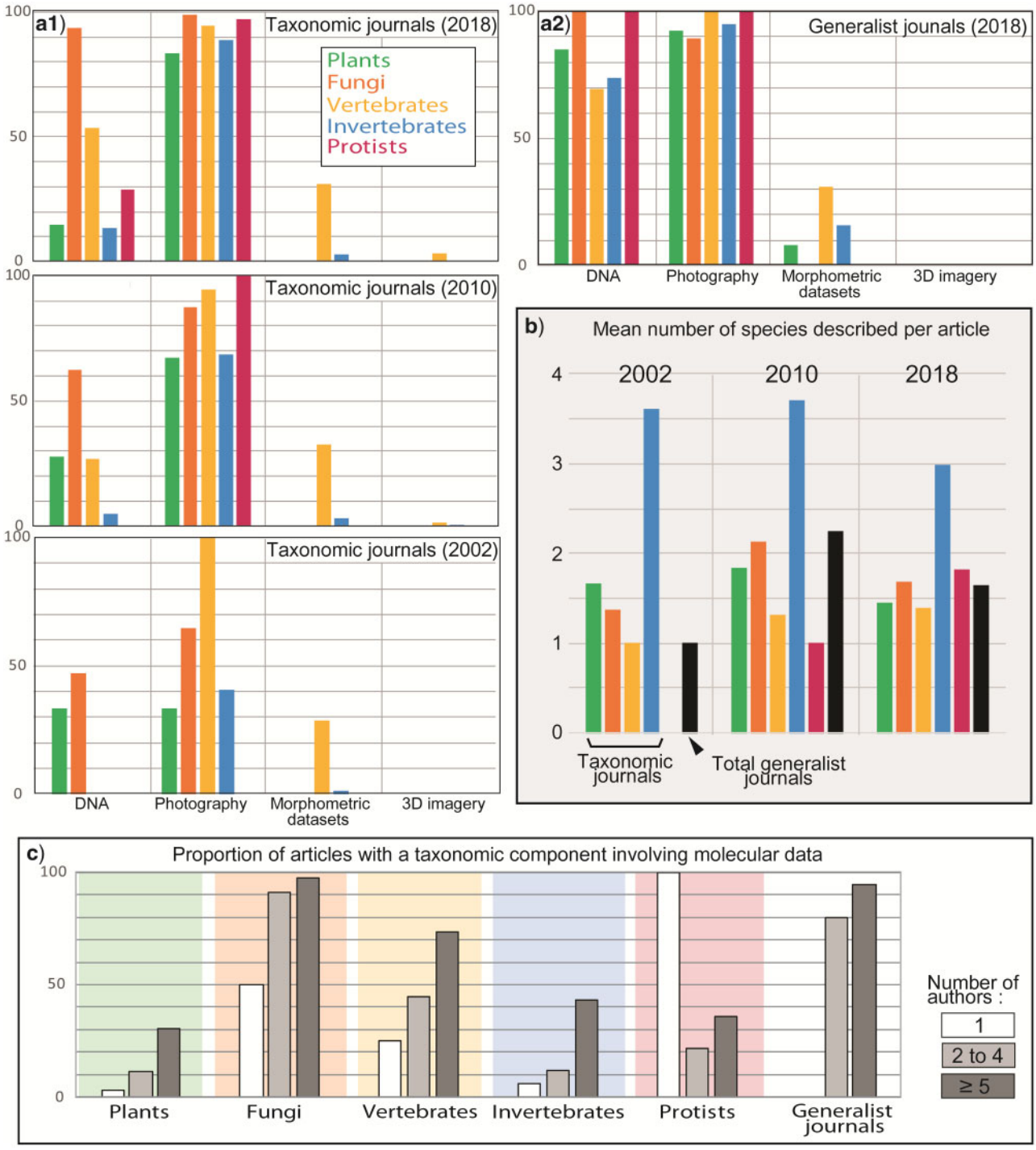

FIGURE 7. a1) Histograms indicating the proportion of alpha-taxonomic studies that have implicated different categories of data in specialized taxonomic journals in 2002, 2010, and 2018. Each series of bars in a and b corresponds to values (from left to right) of plants, fungi, vertebrates, invertebrates, and protists. a2) Same statistics for generalist journals (only 2018 is represented for these journals, the number of papers dealing with alpha-taxonomic issues being negligible for 2002 and 2010, with only 3 and 5 new species during these 2 years). (b) Mean number of species named per article in 2002, 2010, and 2018 in specialized taxonomic journals and generalist journals. c) Proportion of articles with a taxonomic component involving molecular data as a function of the number of authors. Specialized taxonomic journals are represented by a selection of four journals with a strong taxonomic component: *Mycological Progress* (Mycology), *Phytotaxa* and *Systematic Botany* (Botany), and *Zootaxa* (Zoology). Taxonomic works dealing with protists are shared among these journals. These journals belong to the top journals with taxonomic orientation and were selected according to our subjective opinion. The generalist journal category includes *PLoS ONE*, *Scientific Reports*, *Nature*, *Science*, *Biological*, *Zoological* and *Botanical Journals the Linnean Society*, and *PNAS*. "DNA" refers to mitochondrial or nuclear sequence data sets, "Photography" to classical photography plus pictures generated by light and electron microscopy, "Morphometry" to all sets of measurements realized and reported with a comparative perspective on a set of several specimens, and "3D imagery" to every study that generated data using tomographic methods (mostly 3D X-ray μCT, plus one paper using synchrotron radiation μCT). See details in Supplementary Appendices S6–S8 available on Dryad.

Of other kinds of data, measurement-based morphometric analysis was used relatively frequently (348 studies), whereas landmark-based 2D- or 3D-morphometry was rarely applied (7 studies only); 9 studies used geographical models; 13 reported or analyzed extensive ecological data sets, including variables ranging from climate to culture media; 78 studies used analysis of sounds (of vertebrates and insects); and 5 used analyses of electric waves, vibrations, and similar signals.

Our survey may be biased against innovative and groundbreaking taxonomic discoveries because those are often published in generalist journals of higher-impact factor. The data we obtained from the generalist journals surveyed (Supplementary Appendix S5 available on Dryad) confirmed this suspicion, with 69% of papers on all organismal categories discussing DNA data. Overall, molecular data were rare in papers published by single authors, whereas papers published by larger author teams mentioned such data more frequently (Fig. 7c, Supplementary Appendix S8 available on Dryad). Taxonomists from each of five global regions use similar proportions of the data types (2D imagery > DNA > morphometrics > 3D data; Supplementary Appendices S9 and S10 available on Dryad).

The journals *Zookeys* and *Phytokeys*, established in 2008 and 2014 respectively, and hence not included in our main survey, encourage data sharing and automatic linking of metadata, and the aims of *Zookeys* (zookeys.pensoft.net, accessed 22 August 2019) include the "preservation of digital materials to meet the highest possible standards of the cybertaxonomy era." Yet, the general pattern of data use in these two journals so far does not differ from that in other outlets. In 2018, for all 83 alpha-taxonomic papers published in *Phytokeys*, and 100 randomly chosen ones published in *Zookeys*, molecular data were implicated in 29% (botany), 22% (entomology), and 50% (vertebrates). Despite innovations such as semantic markup or tagging, a method that assigns markers, or tags, to taxonomic names, gene sequences, localities, designations of nomenclatural novelties, and so on (Penev et al. 2018), standardization and sharing of raw data are far from being widely implemented in taxonomy. For instance, only 2.5% of all the GBIF-mediated occurrences for the 24 classes of organisms surveyed by Troudet et al. (2018) were linked to digital data and 1.5% to DNA sequences, and outlets such as the *Biodiversity Data Journal* (Smith et al. 2013) that try to redefine taxonomic papers as sources of data rather than narratives, remain an exception—probably not only because of technological limitations but also motivational factors (Hipsley and Sherratt 2019).

How many DNA sequences are produced in the context of taxonomic research? We used *Zootaxa* as a benchmark, representative of a large amount of contemporary taxonomic work. For 2015–2018, numbers of sequences deposited in NCBI-GenBank (accessed August 22, 2019) with a *Zootaxa* reference varied between 8662 and 14,073 per year (Supplementary Appendix S11 available on Dryad). With 2321 papers published in the journal in 2018, this corresponds to an average of six DNA sequences per taxonomic study. While this may be an underestimate because taxonomists often report the results of their molecular phylogenetic studies separately in higher-impact journals, the overall picture is that taxonomy is not yet fully embracing the opportunities offered by the analysis of genetic data.

Our analysis indicates that images are the most universal data type produced in alpha-taxonomic work. This is true of all regions of the world (Supplementary Appendix S9 available on Dryad). As a conservative estimate, 10 images may typically be produced of the holotype and paratypes of a new species and published as part of the taxonomic study. Mostly, these are photographs and drawings, sometimes scanning electron microscopy (SEM). We may assume that in comprehensive revisionary studies, up to 100 images (of comparative voucher specimens, or of different morphological characters) will be produced per newly named species. Most are probably neither published nor submitted to repositories. Assuming again 20,000 new species named per year (Fig. 1), and a bound of 100 images per new species, this leads to an estimated $\leq$2 million images produced per year in the context of alpha-taxonomic studies. Considering that Instagram alone hosts more than 50 billion images and accepts more than 100 million new images per day (www.omnicoreagency.com, accessed January 19, 2020), the yearly storage capacity required for taxonomy-specific images produced in alpha-taxonomic research appears manageable and in the short term is smaller than that needed for intensive digitization campaigns of natural history museums and herbaria (e.g., Le Bras et al. 2017).

Useful Data for Next-Generation Taxonomy

Our survey revealed that taxonomists in their routine alpha-taxonomic work do not make systematic use of large omics data sets or 3D imagery. A rise in the use of such advanced molecular and imagery data sets, however, is likely, especially as these methods become more affordable and as images of the type specimens of new names may become required by the codes of nomenclature. Taxonomists' requirements for data and metadata formats, however, go beyond DNA sequences and images. Verifiability of taxonomic work may sometimes require the archiving of computer memory-intensive raw data of genomic and transcriptomic studies, for example, in the NCBI-SRA Sequence Read Archive, but assemblies, especially if findable via a specimen identifier and accompanied by specimen metadata, may be more important. So far, however, assemblies especially of RNAseq experiments are often not submitted to the Transcriptome Shotgun Assembly Sequence Database (https://www.ncbi.nlm.nih.gov/genbank/tsa/) or other specialized repositories in a searchable format (Moreton et al. 2015).

Geographical occurrence data, also extremely important for taxonomic work, are available from GBIF (https://www.gbif.org/; 1.3 billion records as of September 2019) or Map of Life (https://mol.org/) and furthered also by citizen science portals (e.g., iNaturalist, https://www.inaturalist.org/), but metabarcoding data, which include occurrence records of morphologically cryptic or microscopic taxa including fungi, protists, or small invertebrates, are so far not stored in a retrievable way. This is because the focus has been on archiving the raw sequence reads rather than the consensus OTU sequences that could be reused by taxonomists. Standards for metabarcoding data should therefore include the archiving of quality-filtered consensus reads in a searchable format, preferably as species hypotheses linked to DOI numbers (Tedersoo et al. 2015).

Lastly, chemotaxonomy is routine in the taxonomy of prokaryotes (Stackebrandt and Smith 2019), is often used in fungi (Frisvad et al. 2008), has proven useful in several classification approaches in plants (Wink et al. 2010), and may be useful for some insects (Kather and Martin 2012) and vertebrates (Poth et al. 2012; Starnberger et al. 2013). According to our survey, it is rarely used in alpha-taxonomic studies of nonfungal eukaryotes today, but metabolomic or proteomic profiles (Steinmann et al. 2013; Rossel and Martínez 2019) and NIR spectra (Rodríguez-Fernández et al. 2011; Kinzner et al. 2015) have proven useful in large-scale species identification and discrimination. Chemotaxonomic data traditionally play an important role in lichenized fungi (Lumbsch 2002), and mycologists distinguish species by HPLC profiling (Kuhnert et al. 2017; Helaly et al. 2018) and sometimes higher taxa based on secondary metabolites (Wendt et al. 2018). The retention factors of known chemotaxonomic markers in standard thin-layer chromatography systems are stored in the LIAS database (http://www.lias.net/). For spectroscopic data including GC-MS, the NIST database (https://www.nist.gov/pml/atomic-spectra-database) provides reference spectra for many plant metabolites but does not act as a repository. Chemotaxonomy can be aided by commercial databases like DNP, (http://dnp.chemnetbase.com/), which contains comprehensive information about the occurrence and distribution of secondary metabolites across all organism kingdoms but these databases are not open access and incur considerable license fees. Metabolomic and chemotaxonomic repositories do exist (e.g., Tsugawa et al. 2019) but the underlying raw data may vary in quality and quantity depending on the applied technological sensitivity, and thus may not be readily searchable or comparable across platforms.

Criteria for Taxonomic Data Repositories

The importance of data repositories becoming part of the routine taxonomic research workflow was recognized almost 20 years ago (Louis et al. 2002; Lynch 2008). Today, there is a plethora of repositories, many of them highly specialized (Louis et al. 2002; Pampel et al. 2013). Of the few generalist repositories, some are not free of charge, and many do not provide curated metadata that would allow informed searches (Assante et al. 2016). Many scientific journals in the life sciences now recommend data repositories for archiving the data that accompany a paper (e.g., the journal Scientific Data on behalf of *Springer Nature* journals: https://www.nature.com/sdata/policies/repositories, or *PLoS*: Public Library of Science Recommended Data Repositories; DOI: 10.25504/FAIRsharing.t2exm). Dedicated registries have been developed to searching repositories for specific kinds of data (e.g., re3data.org/ and fairsharing.org/), with the FAIR data principles—data should be *Findable, Accessible, Interoperable*, and *Reusable*—as a framework (Wilkinson et al. 2016) and measurable metric (Wilkinson et al. 2018). Taxonomic data repositories should be (i) free of charge for data contributors, (ii) user-friendly, with a low-complexity submission workflow, not requiring affiliation to academic institutions and not requiring cumbersome registration or login procedures, and (iii) including careful and prompt quality-checks of submissions by dedicated data curators. This is particularly important because a substantial proportion of the estimated 30,000–40,000 taxonomists worldwide (Haas and Häuser 2007) lack data management expertise and support as they often work as single authors or small teams (Knapp 2008; Joppa et al. 2011) and in many cases are nonprofessional researchers (Hopkins and Freckleton 2002; Fontaine et al. 2012).

Ideally, taxonomic repositories should be able to handle universally unique identifiers to refer to specimens (Guralnick et al. 2015; Güntsch et al. 2018; Nelson et al. 2018; Triebel et al. 2018). At present, however, a mandatory use of such identifiers for submission of taxonomic data is unrealistic because, as we have explained above, (i) they do not yet exist for many collections and (ii) the best way of numbering bulk collections is still unclear. For data reuse to be encouraged and facilitated in taxonomy and by its end users, emphasis should be on making data and metadata available in highly standardized formats, enhancing comparability across taxonomic studies. Metadata should thus include a specimen identifier in best-practice format for the respective group of organisms, in addition to a species-level name (accepted or candidate species) and information on geographic location, if possible including geographical coordinates. Usage of standards defined in the Darwin Core or ABCD (Holetschek et al. 2012; Wieczorek et al. 2012) would be highly advisable. In general, however, the submission procedure should keep mandatory metadata to a minimum but provide an extensive, standardized list of optional metadata, as in the minimum checklist concept of the Minimum Information about any (x) Sequence (MIxS) for DNA data (Yilmaz et al. 2011).

Taxonomy is firmly grounded in history. Studies published 100 or 200 years ago are regularly consulted

by taxonomists today and so are voucher specimens collected over centuries (see also Venu and Sanjappa 2011). The principal task of natural history museums and herbaria is to preserve biological materials in perpetuity. The rapid technological turnover of the digital era therefore elicits concerns in the taxonomic community (e.g. Dubois 2003; Padial and De la Riva 2007): can data storage be ensured for "perpetuity"? This concern may be alleviated by data repositories acquiring a certificate, like the CoreTrustSeal (https://www.coretrustseal.org/), which certifies that they are sustainable and trustworthy. Because museums and herbaria already provide long-term storage and careful curation of specimens, their data centers are also the ideal location for long-term repositories of specimen-associated data, certified under even stricter rules such as requiring a well-defined exit strategy defining where the data will be archived if the repository ceases to exist (Table 3).

Taxonomic data repositories should include (i) the option of complex advanced searches with elaborate combinations of inclusion and exclusion of search terms (and/or an API), (ii) semantic (contextual) searches for finding species under synonymous names, (iii) fuzzy searches allowing for different spelling variations e.g. of specimen identifiers, and (iv) the option to search a repository through other, general portals like GBIF (gbif.org) or GFBio (gfbio.org). Searches that include taxon names could be facilitated by the possibility to access established taxonomic backbones, such as the NCBI taxonomy (Federhen 2012), GBIF, or the many databases underlying the Catalogue of Life (http://www.catalogueoflife.org/), or ideally to a dynamic database providing a Global Names Architecture (Pyle 2016).

Large-scale taxonomic studies are often impeded by the sheer amount of data that need to be compared. The problem is compounded by an inherent conflict between the two main interests of taxonomy—quality and speed of delimitation (Sangster and Luksenburg 2015). Probabilistic tools for (semi-)automated species delimitation relying on high-quality data repositories might help. A few such tools have been developed, including Structure (Pritchard et al. 2000), GMYC (Pons et al. 2006), Haploweb (Flot et al. 2010), ABC (Camargo et al. 2012), ABGD (Puillandre et al. 2012), RESL (Ratnasingham and Hebert 2013), and PTP (Zhang et al. 2013), but they all rely on DNA data and do not integrate other taxonomic evidence (Edwards and Knowles 2014). Examples of programs for automated integrative species delimitation (including information from geography or morphology) are Geneland (Guillot et al. 2005) and iBPP (Solís-Lemus et al. 2015). In the future, initial species delimitation hypotheses could be elaborated by probabilistic (machine-learning) algorithms that make full use of data from different repositories. For this to work, data in repositories need to be machine-accessible, standardized, reviewed, georeferenced, and current.

A final criterion for taxonomic data repositories is flexibility in format because of the diversity of taxonomic data (above and Figs. 3 and 4). To reflect this diversity, data submission should allow for user-defined metadata formats, but enforce the use of Darwin Core or ABCD standards (Holetschek et al. 2012; Wieczorek et al. 2012; Cicero et al. 2017) where applicable and not impose restrictions on the number of data files to be submitted. None of the 15 taxonomic repositories reviewed for this article meet all 12 of the needs and criteria assessed (Tables 3 and 4, Supplementary Appendix S12 available on Dryad). Some criteria, especially free and open access, are fulfilled by most repositories, but taxonomy-specific options for submission or search are not. As examples, the leading repositories in the field of molecular data (GenBank, http://www.ncbi.nlm.nih.gov/genbank; DDBJ, https://www.ddbj.nig.ac.jp; ENA, https://www.ebi.ac.uk/ena) seem to be compliant with most of the criteria in Table 3. In contrast, taxonomy-specific repositories, for instance those for bioacoustic recordings in amphibian taxonomy (Köhler et al. 2017), do not make data openly available for reuse.

Recommendations and Conclusions

The last decades have seen a massive increase of taxonomic cyber-infrastructure, delivering crucial services to many end users. Only a minor fraction of this infrastructure has, however, been specifically conceived to support the alpha-taxonomic workflow itself. Taxonomists themselves need to become more involved with the development of tools to integrate the existing resources into their operational pipelines. Perhaps most important are data portals to retrieve and submit specimen-based data. Via customized searches, a taxonomic portal fully dedicated to aggregating data based on specimen identifiers would retrieve all data in real time—DNA sequences, images, current species attribution—available for a specimen across distributed repositories and databases, thus coming close to the cyberspecimen concept. Distributed collection catalog portals, in particular VertNet (http://vertnet.org/), already have implemented many of the search options needed by taxonomists and could be successively expanded (Cicero et al. 2017). Connecting such a catalog to molecular data repositories, especially GenBank (https://www.ncbi.nlm.nih.gov/genbank/) or the Barcode of Life (http://www.boldsystems.org/), whose structure fits our criteria for taxonomic data repositories quite well (Table 4) seems to be a logical first step. Repositories should also be linked with taxonomic databases in a flexible way, allowing data to be retrieved not only under the current taxonomic name but also in nomenclatural and perhaps taxonomic synonym searches. A closer collaboration of taxonomists with the data scientists working on large cybertaxonomy projects in the same institutions may create unexpected synergies

TABLE 3. Criteria relevant for specimen-based taxonomic data repositories.

Priority	Criterion	Explanation
1	Specimen-based data structure	As alpha-taxonomy is centered on specimens, the repository structure must allow for the identification of data from specimen numbers. Both submission and retrieval/search must include a specimen identifier option.
2	Sustainability—certainty of perpetual data storage	The naming of organisms is based on the principle of historical priority, and in taxonomy, publications and data do not lose importance over time. The long-term availability of taxonomic data is therefore a *sine qua non* condition for repositories. This include, but is not limited to, long-term funding (preferably permanent), adequate data backups and if possible, existence of mirrors and contingency strategies.
3	Adherence to the FAIR principles	The principles of findable, accessible, interoperable, and reusable are partly overlap with the more specific conditions listed in this table; still, overall adherence to the FAIR principles constitutes an important criterion, measurable by "Fair Metrics" (Wilkinson et al. 2018).
4	Free of charge for data submitters and open access for data users	Many taxonomists do not have access to institutional funds, and many taxonomic journals do not cover repository fees. To be successful in capturing an increasing proportion of taxonomy-related data, a repository must not charge data submission fees.
5	User-friendly low-complexity workflow for data submission	Time-consuming submission procedures act as strong deterrent in convincing the large community of taxonomists (including amateurs) of the value of making their data available. Furthermore, given the enormous differences among collections in defining and labeling specimens, data-deficient historical specimens, and nonstandardized collections across the world, the amount of mandatory data fields for submission should be minimal (specimen identifier, species name, geographic location).
6	Submission and storage of data packages from multicollection sets of specimens	Taxonomists typically revise a group of organisms by examining specimens from collections held by multiple institutions, often from different countries and continents. Repositories should allow for coherent data packages containing such multicollection data rather than institutional or national repositories restricting data to those from their collection or country.
7	Data submission portal with options for taxonomic (specimen-based) data	Even if a repository allows for specimen identifiers, the submission tools are often not optimized for taxonomy-related data. Ideally, a repository should allow bulk submissions of many kinds of data (e.g., DNA sequences, images), linked to specimen identifiers by a separate metadata table.
8	Machine-accessible for automated data retrieval	Given the prospect of machine-learning tools for species delimitation and species identification, the information in a repository should be automatically retrievable and readable through the web.
9	Link to taxonomic databases for species identifiers, synonymies, etc.	The assignment of species names to taxonomic data is secondary because these names are bound to change over time. Yet, to facilitate their retrieval, data should be associated as much as possible with accepted and valid genus and species names. Through dynamic links to taxonomic databases, entries can be assigned to species names even if originally entered under different synonyms, declensions, or combinations.
10	Compliance with taxonomic data standards	While allowing for flexibility and enforcing only a minimal number of metadata fields per data item is preferable, repositories for taxonomic data should ideally be structured in agreement with international taxonomy standards: metadata field names should agree with Darwin Core or ABCD terminology, specimen identifiers should allow for CETAF standards.
11	Manual search options tailored to the needs of taxonomists	To reflect variation of taxonomic questions, advanced, semantic, and fuzzy searches are desirable.
12	Data searches possible through other portals	Repositories should be favored for taxonomic data if they are linked to overarching data portals which can be used to search multiple repositories at once.
13	No limitation to number of data files	Since data packages for taxonomic monographs may contain data on hundreds or thousands of specimens, a repository should not enforce an a priori limit on the number of data items per submission.
14	Wide use and acceptance by the community	Reinventing the wheel should be avoided and repositories widely accepted and used by the community should be preferred, i.e., repositories (i) where many data have already been submitted by (ii) a large number of different submitters, and (iii) which are listed as standard and recommended by journals and publishers (e.g., *Springer Nature* and *PLoS* lists).

because often, small modifications to existing data-aggregating portals could substantially improve their utility for taxonomists.

Images are among the most widely produced and used types of data in alpha-taxonomy (Fig. 7). Establishing portals that allow image repositories to be searched by specimen identifiers should become a priority. Images are semistructured data, and successful managing or searching of such data requires metadata, including species identifiers, annotations, scale information, authorship, and geographical location. New software solutions are needed to collect and safeguard this information and the diverse image data. Recently, image annotation software tools have been proposed to support, for instance, environmental monitoring (Schlining and Stout 2006; Kloster et al. 2014; Althaus et al. 2015; Beijbom et al. 2015; Langenkämper et al. 2017). These tools are easy to use and have low requirements of computational power (Zurowietz et al. 2019). Most of them are already equipped with machine-learning functions to automate some steps in the annotation process. Toolboxes to be included in taxonomic repositories, or in cyberspecimen data portals, could include automatic detection of rulers or scale bars, dynamic continuous zoom, and measurement tools both for 2D and 3D images.

Versatile data portals connected to rich taxonomic data repositories would benefit taxonomists as well as end users of taxonomy. For instance, the progress in computational power and imaging technology on smartphones allows the collection of visual data and the instant availability of taxonomic knowledge on a new scale. There is a boom of cellphone apps that identify species of plants and mushrooms (e.g., Pl@ntNet, https://identify.plantnet.org/; PlantSnap, https://www.plantsnap.com/; Naturblick, http://www.naturblick.naturkundemuseum.berlin) or

TABLE 4. Evaluation of a selection of 15 repositories according to 12 of the 14 criteria listed in Table 3 (++= compliant; += unsatisfactory or partial; −= not compliant)

Criterion	1. Specimen-based structure [1.1.submission]	[1.2. search]	2. Sustainability	3. Compliance with FAIR criteria	4. Free of charge [4.1. submitters]	[4.2. users]	6. Multicollection data package	7. Submission portal with taxonomic options	9. Link to taxonomic databases	10. Compliance with taxonomic data standards	11. Taxonomic search options	12. Access through other portals	13. No limitation to number of data	14. Wide use/acceptance by the community
DRYAD	+	−	++	++	−	++	++	−	−	+	+	−	+	++
Figshare	+	+	+	++	++	++	++	−	−	+	−	−	+	+
Macaulay Library	++	++	+	−	++	+	++	+	+	++	−	++	++	++
PANGEA	+	−	++	++	++	++	++	+	−	++	−	++	++	++
OSF	−	−	++	++	++	++	++	−	−	−	−	−	++	+
Morphobank	+	+	+	++	++	++	++	−	−	++	−	−	++	++
Digimorph	++	+	−	−	++	+	++	−	−	−	−	−	+	+
Morphomuseum	++	++	−	−	++	++	++	−	−	++	+	−	++	+
Morphosource	++	++	++	++	++	++	++	+	−	+	+	−	++	++
IDR	+	+	+	++	++	++	++	++	−	−	−	+	++	+
Metabolights	++	++	+	++	++	++	++	+	−	+	−	−	++	++
Genbank	++	++	++	++	++	++	++	+	++	++	++	+	++	++
DDBJ	++	++	++	++	++	++	++	+	++	++	++	+	++	++
ENA	++	++	++	++	++	++	++	+	++	++	++	++	++	++
Movebank	++	−	+	++	++	++	++	++	−	+	−	++	++	+

See details in Supplementary Appendix S12 available on Dryad.

animals (e.g., https://fieldguide.ai/) or all of the above (https://www.inaturalist.org) by automated comparison of photos with large image collections. Similar apps also exist for sound-based species identification of birds (e.g., SongSleuth, https://www.songsleuth.com/; BirdNet, https://birdnet.cornell.edu/; BirdGenie, https://press.princeton.edu/apps/birdgenie.html; BirdSongID, http://isoperla.co.uk/; ChirpOMatic, http://www.chirpomatic.com/), bats (e.g. iBatsID, https://sites.google.com/site/ibatsresources/iBatsID), and increasingly also insects (e.g. CicadaHunt, http://newforestcicada.info/app/). These apps impressively demonstrate the potential of computer-based approaches to species identification and provide a glimpse into what may be possible in a future in which large virtual collections of cyberspecimens become available to train artificial intelligence pipelines.

Having reviewed numerous data repositories for this study, we propose a pilot submission template in Supplementary Appendices S13 and S14 available on Dryad, building upon models established by the NCBI Sequence Read Archive and (re-)using ABCD terms. This template is currently being tested for the submission of data to the GFBio data centers (Diepenbroek et al. 2014). Because taxonomy is intrinsically dependent on long-term availability of data, taxonomists will have a high motivation to meet the "taxonomic data repository" challenge and to develop concepts of truly sustainable, potentially perpetual data storage. The electricity usage and the carbon footprint associated with data storage (Andrae and Edler 2015; Jones 2018) may require standards allowing submitters to identify which data truly merit long-term storage (e.g., to prevent submission of redundant or blurred pictures, or to optimize their resolution level when it is excessively high). A stringent archiving strategy of original taxonomic data could become an integral part of a renewed procedure to name new species—accelerated but without compromising quality of species hypotheses, mobilizing species information through images, DNA sequences, sounds, or tabulated trait information, while relieving taxonomists from manually compiling lengthy descriptions. Although words will necessarily remain the means to justify taxonomic decisions, evaluate species criteria and (briefly) list diagnostic features of new species, taxonomists should consider moving towards publishing alpha-taxonomic results as interlinked, standardized, and openly accessible data sets rather than traditional descriptive papers.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.fj6q573qd.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, grant number DFG RE 603/29-1) and benefited from the sharing of expertise within the DFG priority program SPP 1991 Taxon-Omics.

Acknowledgments

We are grateful to William N. Eschmeyer, Jon D. Fong, Ronald Fricke, Darrel R. Frost, Rafaël Govaerts, Vincent Robert, Peter Uetz, and Richard van der Laan for useful advice and data on rates of species discovery and naming. We thank Christy Hipsley and one anonymous reviewer for constructive feedback on our manuscript. We also thank Steve A. Marshall, Neal Evenhuis, and Sébastien Soubzmaigne for allowing the use of original photographs.

References

Akkari N., Enghoff H., Metscher B.D. 2015. A new dimension in documenting new species: high-detail imaging for myriapod taxonomy and first 3D cybertype of a new millipede species (Diplopoda, Julida, Julidae). PLoS One 10:e0135243.

Althaus F., Hill N., Ferrari R., Edwards L., Przeslawski R., Schönberg C. H., Stuart-Smith R., Barrett N., Edgar G., Colquhoun J., Tran M., Jordan A., Rees T., Gowlett-Holmes K. 2015. A standardised vocabulary for identifying benthic biota and substrata from underwater imagery: the catami classification scheme. PLoS One 10:e0141039.

Amorim D.S., Santos C.M.D., Krell F.T., Dubois A. 2016. Timeless standards for species delimitation. Zootaxa 4137:121–128.

Andrae A.S.G., Edler T. 2015. On global electricity usage of communication technology: trends to 2030. Challenges 6:117–157.

Anonymous [International Commission on Zoological Nomenclature]. 1999. International code of zoological nomenclature. 4th ed. London: International Trust for Zoological Nomenclature, p. i–xxix + 1–306.

Assante M., Candela L., Castelli D., Tani A. 2016. Are scientific data repositories coping with research data publishing? Data Sci. J. 15:6.

Balke M., Schmidt S., Hausmann A., Toussaint E.F.A., Bergsten J., Buffington M., Häuser C.L., Kroupa A., Hagedorn G., Riedel A., Polaszek A., Ubaidillah R., Krogmann L., Zwick A., Fikáèek M., Hájek J., Michat J.C., Dietrich C., La Salle J., Mantle B.K.L., Ng P., Hobern D. 2013. Biodiversity into your hands—a call for a virtual global natural history 'metacollection'. Front. Zool. 10:55.

Beijbom O., Edmunds P., Roelfsema C., Smith J., Kline D., Neal B., Dunlap M.J., Moriarty V., Fan T.Y., Tan C.J., Chan S., Treibitz T., Gamst A., Mitchell B.G., Kriegman D. 2015. Towards automated annotation of benthic survey images: variability of human experts and operational modes of automation. PLoS One 10:e0130312.

Bik H.M. 2017. Let's rise up to unite taxonomy and technology. PLoS Biol. 15:e2002231.

Bosselaers J., Dierick M., Cnudde V., Masschaele B., van Hoorebeke L., Jacobs P. 2010. High-resolution X-ray computed tomography of an extant new Donuea (Araneae: Liocranidae) species in Madagascan copal. Zootaxa 2427:25–35.

Brooke M. de L. 2000. Why museums matter. Trends Ecol. Evol. 15:136–137.

Camargo A., Morando M., Avila L.J. and Sites J.W. 2012. Species delimitations with ABC and other coalescent-based methods: a test of accuracy with simulations and an empirical example with lizards of the *Liolaemus darwinii* complex (Squamata: Liolaemidae). Evolution 66:2834–2849.

Cannon P., Aguirre-Hudson B., Aime M.C., Ainsworth A.M., Bidartondo M.I., Gaya E., Hawksworth D., Kirk P., Leitch I.J., Lücking R. 2018. Definition and diversity. In: Willis K.J., editors. State of the world's fungi. Report. Kew: Royal Botanic Gardens, p. 4–11.

Ceriaco L.M.P., Gutiérrez E.E., Dubois, A. 2016. Photography-based taxonomy is inadequate, unnecessary, and potentially harmful for biological sciences. Zootaxa 4196(3): 435–445.

Chauvel B., Dessaint F., Cardinal-Legrand C., Bretagnolle, F. 2006. The historical spread of *Ambrosia artemisiifolia* L. in France from herbarium records. J. Biogeogr. 33:665–673.
Christidis L. (Ed.) 2018. The Howard and Moore complete checklist of the birds of the world, version 4.1 (Downloadable checklist). Available from: https://www.howardandmoore.org (March 15, 2019).
Cicero C., Spencer C.L. Bloom D.A., Guralnick R.P., Koo M.S., Otegui J.. Russell L.A, Wieczorek J.R. 2017. Biodiversity informatics and data quality on a global scale. In: Webster M.S., editors. Emerging frontiers in collections-based ornithological research: the extended specimen. Studies in avian biology. Boca Raton, FL: CRC Press, p. 201–218.
Costello M.J., May R.M., Stork N.E. 2013a. Can we name Earth's species before they go extinct? Science 339(6118):413–416.
Costello M.J., Wilson S., Houlding B. 2013b. More taxonomists describing significantly fewer species per unit effort may indicate that most species have been discovered. Syst. Biol. 62:616–624.
Crous P.W., Gams W., Stalpers J.A., Robert V., Stegehuis G. 2004. MycoBank: an online initiative to launch mycology into the 21st century. Stud. Mycol. 50(1):19–22.
De Mauro A., Greco M., Grimaldi M. 2016. A formal definition of big data based on its essential features. Library Rev. 65:122–135
de Queiroz K. 1998. The general lineage concept of species, species criteria, and the process of speciation. In: Howard D.J., Berlocher S.H., editors. Endless forms: species and speciation. New York: Oxford University Press., p. 57–75.
de Queiroz K. 2007. Species concepts and species delimitation. Syst. Biol. 56:879–886.
Diepenbroek M., Glöckner F., Grobe P., Güntsch A., Huber R., König-Ries B., Kostadinov I., Nieschulze J., Seeger B., Tolksdorf R., Triebel D. 2014. Towards an integrated biodiversity and ecological research data management and archiving platform: the German Federation for the Curation of Biological Data (GFBio) In: Plödereder E., Grunske L., Schneider E., Ull D., editors. Informatik 2014—big data komplexität meistern. GI-Edition: Lecture Notes in Informatics (LNI)—Proceedings. GI edn., vol. 232. Bonn: Köllen, p. 1711–1724.
Dietrich C., Hart J., Raila, D., Ravaioli U., Sobh N., Sobh O., Taylor C. 2012. InvertNet: a new paradigm for digital access to invertebrate collections. Zookeys 209:165–181.
Dubois A. 2003. Should internet sites be mentioned in the bibliographies of scientific publications? Alytes 21:1–2.
Edwards D.L., Knowles L.L. 2014. Species detection and individual assignment in species delimitation: can integrative data increase efficacy? Proc. R. Soc. Lond. [Biol]. 281:20132765.
Faulwetter S., Vasileiadou A., Kouratoras M., Dailianis T., Arvanitidis C. 2013. Micro-computed tomography: introducing new dimensions to taxonomy. Zookeys 263:1–45.
Favret C. 2014. Cybertaxonomy to accomplish big things in aphid systematics. Insect Sci. 21:392–399.
Federhen S. 2012. The NCBI taxonomy database. Nucleic Acids Res. 40 (Database issue):D136–D143.
Flot J.-F., Couloux A., Tillier S. 2010. Haplowebs as a graphical tool for delimiting species: a revival of Doyle's "field for recombination" approach and its application to the coral genus *Pocillopora* in Clipperton. BMC Evol. Biol. 10:1–14.
Fontaine B., van Achterberg K., Alonso-Zarazaga M.A., Araujo R., Asche M., Aspöck H., Aspöck U., Audisio P., Aukema B., Bailly N., Balsamo M., Bank R.A., Belfiore C., Bogdanowicz W., Boxshall G., Burckhardt D., Chylarecki P., Deharveng L., Dubois A., Enghoff H., Fochetti R., Fontaine C., Gargominy O., Gomez Lopez M.S., Goujet D., Harvey M.S., Heller K.G., van Helsdingen P., Hoch H., De Jong Y., Karsholt O., Los W., Magowski W., Massard J.A., McInnes S.J., Mendes L.F., Mey E., Michelsen V., Minelli A., Nieto Nafrıa J.M., van Nieukerken E.J., Pape T., De Prins W., Ramos M., Ricci C., Roselaar C., Rota E., Segers H., Timm T., van Tol J., Bouchet P. 2012. New species in the old world: Europe as a frontier in biodiversity exploration, a test bed for 21st century taxonomy. PLoS One 7:e36881.
Frisvad J.C., Andersen B., Thrane U. 2008. The use of secondary metabolite profiling in chemotaxonomy of filamentous fungi. Mycol. Res. 112(2):231–240.
Frost D.R. 2019. Amphibian species of the world: an online reference. Version 6.0. Website. Available from: http://research.amnh.org/herpetology/amphibia/index.html. American Museum of Natural History, New York, USA (March 15, 2019).
Garraffoni A.R.S., Freitas A.V.L. 2017. Photos belong in the taxonomic code. Science 355(6327):805.
Gemeinholzer B., Vences M., Beszteri B., Bruy T., Felden J., Kostadinov I., Miralles A., Nattkemper T.W., Printzen C., Renz J., Rybalka N., Schuster T., Weibulat T., Wilke T., Renner S.S. 2020. Data storage and data re-use in taxonomy—the need for improved storage and accessibility of heterogeneous data. Org. Divers. Evol. 20:1–8.
Gignac P.M., Kley N.J., Clarke J.A., Colbert M.W., Morhardt A.C., Cerio D., Cost I.N., Cox P.G., Daza J.D., Early C.M., Echols M.S., Henkelman R.M., Herdina A.N., Holliday C.M., Li Z., Mahlow K., Merchant S., Müller J., Orsbon C.P., Paluh D.J., Thies M.L., Tsai H.P., Witmer L.M. 2016. Diffusible iodine-based contrast-enhanced computed tomography (diceCT): an emerging tool for rapid, high-resolution, 3-D imaging of metazoan soft tissues. J. Anat. 228(6):889–909.
Godfray H.C.J.Jr. 2007. Linnaeus in the information age. Nature 446:259–260.
Grass A., Tremetsberger K., Hössinger R., Bernhardt K-G. 2014. Change of species and habitat diversity in the Pannonian Region of Eastern Lower Austria over 170 years: using herbarium records as a witness. Nat. Resour. 5:583–596.
Guillot G., Estoup A., Mourtier F., Cosson, J.F. 2005. A spatial statistical model for landscape genetics. Genetics 170:1261–1280.
Güntsch A., Groom Q., Hyam R., Chagnoux S., Röpert D., Berendsohn W., Casino A., Droege G., Gerritsen W., Holetschek J., Marhold K., Mergen P., Rainer H., Smith V., Triebel D. 2018. Standardised globally unique specimen identifiers. Biodivers. Inf. Sci. Stand. 2:e26658.
Guralnick R.P., Cellinese N., Deck J., Pyle R.L., Kunze J., Penev L., Walls R., Hagedorn G., Agosti D. Wieczorek J., Catapano T., Page R. 2015. Community next steps for making globally unique identifiers work for biocollections data. ZooKeys 494:133–154.
Haas F., Häuser C.L. 2007. How many taxonomists are there? Available from: http://www.senckenberg.uni-frankfurt.de/odes/Haas_Haeuser.pdf.
Hawksworth D.L., Hibbett D.S., Kirk P.M., Lücking R. 2016. Proposals to permit DNA sequence data to serve as types of names of fungi. Taxon 65:899–900.
Helaly S.E., Thongbai B., Stadler M. 2018. Diversity of biologically active secondary metabolites from endophytic and saprotrophic fungi of the ascomycete order Xylariales. Nat. Prod. Rep. 35:992–1014.
Hipsley C.A., Sherratt E. 2019. Psychology, not technology, is our biggest challenge to open digital morphology data. Sci. Data. 6:41.
Holetschek J., Dröge G., Güntsch A., Berendsohn W.G. 2012. The ABCD of primary biodiversity data access, Plant Biosyst. 146:771–779.
Hongsanan S., Xie N., Liu J.K., Dissanayake A., Ekanayaka A.H., Raspé O., Jayawardena R.S., Hyde K.D., Jeewon R., Purahong W., Stadler M., Peršoh D. 2018. Can we use environmental DNA as holotypes? Fungal. Divers. 92:1–30.
Hopkins G.W., Freckleton R.P. 2002. Declines in the numbers of amateur and professional taxonomists: implications for conservation. Anim. Conserv. 5:245–249.
IISE 2011. State of observed species. Tempe, AZ: International Institute for Species Exploration. Available from: http:/species.asu.edu/SOS (March 15, 2019).
Jones N. 2018. How to stop data centres from gobbling up the world's electricity. Nature 561:163–166.
Joppa L.N., Roberts D.L., Pimm S.L. 2011. The population ecology and social behaviour of taxonomists. Trends Ecol. Evol. 26:551–553.
Kather R., Martin S.J. 2012. Cuticular hydrocarbon profiles as a taxonomic tool: advantages, limitations and technical aspects. Physiol. Entomol. 37: 25–32.
Kinzner M.C., Wagner H.C., Peskoller A., Moder K., Dowell F.E., Arthofer W., Schlick-Steiner B.C., Steiner F.M. 2015. A near-infrared spectroscopy routine for unambiguous identification of cryptic ant species. PeerJ. 3:e991.
Kloster M., Kauer G., Beszteri B. 2014. SHERPA: an image segmentation and outline feature extraction tool for diatoms and other objects. BMC Bioinformatics 15:218.

Knapp S. 2008. Taxonomy as a team sport. In: Wheeler Q., editor. The new taxonomy. Systematics Association Special Volume 76. London: CRC Press. p. 33–53.
Köhler J., Jansen M., Rodríguez A., Kok P.J.R., Toledo L.F., Emmrich M., Glaw F., Haddad C.F.B., Rödel M.O., Vences M. 2017. The use of bioacoustics in anuran taxonomy: theory, terminology, methods and recommendations for best practice. Zootaxa 4251:1–124.
Krell F.-T. 2015. ZooBank progress report. Bull. Zool. Nomenclat. 72: 181.
Krell F.-T., Marshall S.A. 2017. New species described from photographs: Yes? No? Sometimes? A fierce debate and a new Declaration of the ICZN. Insect Syst. Divers. 1(1):3–19.
Kuhnert E., Sir E.B., Lambert C., Hyde K.D., Hladki A.I., Romero A.I., Rohde M., Stadler M. 2017. Phylogenetic and chemotaxonomic resolution of the genus *Annulohypoxylon* (Xylariaceae) including four new species. Fungal Divers. 85:1–43.
Langenkämper D., Zurowietz M., Schoening T., Nattkemper T.W. 2017. BIIGLE 2.0—browsing and annotating large marine image collections. Front. Mar. Sci, 4:83.
Larsen B.B., Miller E.C., Rhodes M.K., Wiens, J.J. 2017. Inordinate fondness multiplied and redistributed: the number of species on Earth and the new pie of life. Q. Rev. Biol. 92: 229–265.
LaSalle J., Wheeler Q., Jackway P., Winterton S., Hobern D., Lovell D. 2009. Accelerating taxonomic discovery through automated character extraction. Zootaxa 2217:43–55.
Le Bras G., Pignal M., Jeanson M. L., Muller S., Aupic C., Carré B., Flament G., Gaudeul M., Gonçalves C., Invernón V.R., Jabbour F., Lerat E., Lowry P.P., Offroy B., Pimparé Pérez E., Poncy O., Rouhan G., Haevermans T. 2017. The French Muséum national d'Histoire naturelle vascular plant herbarium collection dataset. Sci. Data 4:170016.
Lendemer J., Thiers B., Monfils A.K., Zaspel J., Ellwood E.R., Bentley A., LeVan K., Bates J., Jennings D., Contreras D., Lagomarsino L., Mabee P., Ford L.S., Guralnick R., Gropp R.E., Revelez M., Cobb N., Seltmann K., Aime M.C. 2020. The extended specimen network: a strategy to enhance US biodiversity collections, promote research and education. BioScience 70(1):23–30.
Leonelli S. 2014. What difference does quantity make? On the epistemology of big data in biology. Big Data Soc. 2014:1–11.
Linnaeus C. 1753. Species plantarum exhibentes plantas rite cognitas ad genera relatas, cum differentiis specificis, nominibus trivialibus, synonymis selectis, locis natalibus, secundum systema sexuale digestas. Holmiæ [Stockholm]: Impensis Laurentii Salvi. 132 p.
Linnaeus C. 1758. Systema naturæ per regna tria naturæ, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis. Tomus I. Editio decima, reformata. Holmiæ [Stockholm]: Impensis Laurentii Salvi. 824 p.
Locey K.J., Lennon J.T. 2016. Scaling laws predict global microbial diversity. Proc. Natl. Acad. Sci. USA 113(21):5970–5975.
Lorieul T., Pearson K.D., Ellwood E.R., Goëau H., Molino J.F., Sweeney P.W., Yost J.M., Sachs J., Mata-Montero E., Nelson G., Soltis P.S., Bonnet P., Joly A. 2019. Toward a large-scale and deep phenological stage annotation of herbarium specimens: case studies from temperate, tropical, and equatorial floras. Appl. Plant Sci. 7(3):e01233.
Louis K.S., Jones L.M., Campbell E.G. 2002. Macroscope: Sharing in Science. Am. Sci. 90:304–307.
Lumbsch H.T. 2002. Analysis of phenolic products in lichens for identification and taxonomy. In: Kranner I.C., Beckett R.P., Varma A.K., editors. Protocols in lichenology. Springer Lab Manuals. Berlin, Heidelberg: Springer. p. 281–295.
Lynch C. 2008. Big data: How do your data grow? Nature 455: 28–29.
Marcial L.H., Hemminger B.M. 2010. Scientific data repositories on the web: an initial survey. J. Assoc. Inf. Sci. Technol. 61(10):2029–2048.
Marshall S.A., Evenhuis N.L. 2015. New species without dead bodies: a case for photo-based descriptions, illustrated by a striking new species of *Marleyimyia* Hesse (Diptera, Bombyliidae) from South Africa. ZooKeys 525:117–127.
May T.W., Redhead S.A., Lombard L., Rossman A.Y. 2018. XI International Mycological Congress: report of Congress action on nomenclature proposals relating to fungi. IMA Fungus 9(2):xxii.
Mayden R.L. 1997. A hierarchy of species concepts: the denouement in the saga of the species problem. In: Claridge M.F., Dawah H.A., Wilson M.R. editors. Species: the units of diversity. London, NY: Chapman & Hall. p. 381–423.
McClellan P.H. 2019. Taxonomic punchlines: metadata in biology. Hist. Biol. https://doi.org/10.1080/08912963.2019.1618293.
Miller-Rushing, A.J., Primack R.B., Primack D., Mukunda S. 2006. Photographs and herbarium specimens as tools to document phenological changes in response to global warming. Am. J. Bot. 93:1667–1674.
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 Biol. 9:e1001127.
Moreton J., Izquierdo A., Emes R.D. 2015. Assembly, assessment, and availability of de novo generated eukaryotic transcriptomes. Front. Genet. 6:361.
Nelson G., Sweeney P., Gilbert E. 2018. Use of globally unique identifiers (GUIDs) to link herbarium specimen records to physical specimens. Appl. Plant Sci. 6:e1027.
Padial J.M., De la Riva I. 2007. Taxonomy, the Cinderella of science, hidden by its evolutionary stepsister. Zootaxa 1577:1–2.
Padial J.M., Miralles A., De la Riva I., Vences M. 2010. The integrative future of taxonomy. Front. Zool. 7:16.
Page R.D.M. 2016. DNA barcoding and taxonomy: dark taxa and dark texts. Philos. Trans. R. Soc. B. 371:20150334.
Pampel H., Vierkant P., Scholze F., Bertelmann R., Kindling M., Klump J., Goebelbecker H.J., Gundlach J., Schirmbacher P., Dierolf U. 2013. Making research data repositories visible: the re3data.org registry. PLoS One 8: e78080.
Patterson D.J., Cooper J., Kirk P.M., Pyle R.L., Remsen D.P. 2010. Names are key to the big new biology. Trends Ecol. Evol. 25:686–691.
Penev L., Agosti D., Georgiev T., Senderov V., Sautter G., Catapano T., Stoev P. 2018. The open biodiversity knowledge management (eco-) system: tools and services for extraction, mobilization, handling and re-use of data from the published literature. Biodiver. Inf. Sci. Stand. 2:e25748.
Pons J., Barraclough T.G., Gomez-Zurita J., Cardoso A., Duran D.P., Hazell S., Kamoun S., Sumlin W.D., Vogler A.P. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst. Biol. 55:595–609.
Poth D., Wollenberg K.C., Vences M., Schulz S. 2012. Volatile amphibian pheromones: macrolides of mantellid frogs from Madagascar. Angew. Chem. Int. Ed. 51:1–5.
Pritchard J.K., Stephens M., Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.
Puillandre N., Lambert A., Brouillet S., Achaz G. 2012. ABGD, Automatic barcode gap discovery for primary species delimitation. Mol. Ecol. 21:1864–1877.
Pyle R.L. 2016. Towards a global names architecture: the future of indexing scientific names. Zookeys 550:261–281.
Pyle R.L., Earle J.L., Greene B.D. 2008. Five new species of the damselfish genus *Chromis* (Perciformes: Labroidei: Pomacentridae) from deep coral reefs in the tropical western Pacific. Zootaxa 1671:3–31.
Ratnasingham S., Hebert P.D.N. 2013. A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. PLoS One 8:e66213.
Renner S.S. 2016. A return to Linnaeus's focus on diagnosis, not description: the use of DNA characters in the formal naming of species. Syst. Biol. 65:1085–1095.
Riley J. 2004. Understanding metadata. Bethesda, MD: NISO Press, National Information Standards Organization.
Rissler L.J., Apodaca J.J. 2007. Adding more ecology into species delimitation: ecological niche models and phylogeography help define cryptic species in the black salamander (*Aneides flavipunctatus*). Syst. Biol. 56(6):924–942.
Roch M.A., Batchelor H., Baumann-Pickering S., Berchok C.L., Cholewiak D., Fujioka E., Garland E.C., Herbert S., Hildebrand J.A., Oleson E.M., Van Parijs S., Risch D., Široviæ A., Soldevilla M.S. 2016. Management of acoustic metadata for bioacoustics. Ecol. Inform. 31:122–136.
Roche D.G., Kruuk L.E., Lanfear R., Binning S.A. 2015. Public data archiving in ecology and evolution: how well are we doing? PLoS Biol. 13:e1002295.

Rodríguez-Fernández J.I , De Carvalho C.J.B., Pasquini C., Gomes de Lima K.M, Moura M.O., Carbajal Arizaga, G.G. 2011. Barcoding without DNA? Species identification using near infrared spectroscopy. Zootaxa 2933:46–54.

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

Roskov Y., Ower G., Orrell T., Nicolson D., Bailly N., Kirk P.M., Bourgoin T., DeWalt R.E., Decock W., Nieukerken E. van, Zarucchi J., Penev L., eds. 2019. Species 2000 & ITIS Catalogue of Life, 26th February 2019. Digital resource at www.catalogueoflife.org/col. Species 2000. Naturalis, Leiden, the Netherlands.

Rossel S., Martínez Arbizu P. 2019. Revealing higher than expected diversity of Harpacticoida (Crustacea:Copepoda) in the North Sea using MALDI-TOF MS and molecular barcoding. Sci. Rep. 9:9182.

Rupp K. 2018. 42 Years of microprocessor trend data. Website. Available from: https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/ (March 13, 2019).

Sangster G., Luksenburg, J.A. 2015. Declining rates of species described per taxonomist: Slowdown of progress or a side-effect of improved quality in taxonomy? Syst. Biol. 64:144–151.

Santos C.M.D., Amorim D.S., Klassa B., Fachin D.A., Nihei S.S., Carvalho C.J., Falaschi, R.L., Mello-Patiu C.A., Couri M.S., Oliveira S.S., Silva V.C., Ribeiro G.C., Capellari R.S., Lamas, C.J. 2016. On typeless species and the perils of fast taxonomy. Syst. Entomol. 41:511–515.

Scherz M.D., Glaw F., Vences M., Andreone F., Crottini A. 2016a. Two new species of terrestrial microhylid frogs (Microhylidae: Cophylinae: *Rhombophryne*) from northeastern Madagascar. Salamandra 52:91–106.

Scherz M.D., Ruthensteiner B., Vences M., Glaw F. 2014. A new microhylid frog, genus Rhombophryne, from northeastern Madagascar, and a re-description of *R. serratopalpebrosa* using micro-computed tomography. Zootaxa 3860:547–560.

Scherz M.D., Vences M., Rakotoarison A., Andreone F., Köhler J., Glaw F., Crottini A. 2016b. Reconciling molecular phylogeny, morphological divergence and classification of Madagascan narrow-mouthed frogs (Amphibia: Microhylidae). Mol. Phylogenet. Evol. 100:372–381.

Schlining B.M., Stout, N.J. 2006. "MBARI's Video Annotation and Reference System," OCEANS 2006. Boston, MA: IEEE. p. 1–5.

Short A.E.Z., Dikow T., Moreau C.S. 2018. Entomological collections in the age of big data. Annu. Rev. Entomol. 63:513–530.

Simpson G.G. 1961. Principles of animal taxonomy. New York: Columbia University Press. p. xii + 247.

Small E. 1989. Systematics of biological Systematics (or, Taxonomy of Taxonomy). Taxon 38(3):335–356.

Smith V., Georgiev T., Stoev P, Biserkov J., Miller J., Livermore L., Baker E., Mietchen D., Couvreur T.L., Mueller G., Dikow T., Helgen K.M., Frank J., Agosti D., Roberts D., Penev L. 2013. Beyond dead trees: integrating the scientific process in the Biodiversity Data Journal. Biodivers. Data J. 1:e995.

Solís-Lemus C., Knowles L.L., Ané C. 2015. Bayesian species delimitation combining multiple genes and traits in a unified framework. Evolution 69:492–507.

Stackebrandt E., Smith D. 2019. Paradigm shift in species description: the need to move towards a tabular format. Arch. Microbiol. 201:143–145.

Starnberger I., Poth D., Peram P.S., Schulz S., Vences M., Knudsen J., Barej M.F., Rödel M.-O., Walzl M., Hödl W. 2013. Take time to smell the frogs: vocal sac glands of reed frogs (Anura: Hyperoliidae) contain species-specific chemical cocktails. Biol. J. Linn. Soc. 110:828–838.

Steinmann I.C., Pflüger V., Schaffner F., Mathis A., Kaufmann C. 2013. Evaluation of matrix-assisted laser desorption/ionization time of flight mass spectrometry for the identification of ceratopogonid and culicid larvae. Parasitology 140:318–327.

Stuessy T.F., Crawford D.J., Soltis D.E., Soltis P.S. 2014. Plant systematics—the origin, interpretation, and ordering of plant biodiversity. In: Regnum Vegetabile, vol. 156. Königstein (Taunus): Koeltz Scientific Books. 425 p.

Tedersoo L., Ramirez K.S., Nilsson R.H., Kaljuvee A., Kõljalg U., Abarenkov K. 2015. Standardizing metadata and taxonomic identification in metabarcoding studies. GigaScience 4:34.

Thorpe S.E. 2017. Is photography-based taxonomy really inadequate, unnecessary, and potentially harmful for biological sciences? A reply to Ceríaco et al. (2016). Zootaxa 4226:449–450.

Triebel D., Reichert W., Bosert S., Feulner M., Osieko Okach D., Slimani A., Rambold G. 2018. A generic workflow for effective sampling of environmental vouchers with UUID assignment and image processing. Database 2018:bax096.

Troudet J., Vignes-Lebbe R., Grandcolas P., Legendre F. 2018. The increasing disconnection of primary biodiversity data from specimens: how does it happen and how to handle it? Syst. Biol. 67:1110–1119.

Tsugawa H., Satoh A., Uchino H., Cajka T., Arita M., Arita M. 2019. Mass spectrometry data repository enhances novel metabolite discoveries with advances in computational metabolomics. Metabolites 9(6): pii: E119.

Venu P., Sanjappa M. 2011. The impact factor and taxonomy. Curr. Sci. 101(11):1397.

Webster M.S. 2017. Emerging frontiers in collections-based ornithological research: the extended specimen. Studies in avian biology. Boca Raton, FL: CRC Press. 240 p.

Wendt L., Sir E.B., Kuhnert E., Heitkämper S., Lambert C., Hladki A.I., Romero A.I., Luangsaard J.J., Srikitikulchai P., Peršoh D., Stadler M. 2018. Resurrection and emendation of the Hypoxylaceae, recognised from a multi-gene genealogy of the Xylariales. Mycol. Prog. 17:115–154.

Wheeler Q.D. 2007. Invertebrate systematics or spineless taxonomy? Zootaxa 1668:11–18.

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

Wheeler Q.D., Bourgoin T., Coddington J., Gostony T., Hamilton A., Larimer R., Plaszek A., Schauff M., Solis M.A. 2012b. Nomenclatural benchmarking: the roles of digital typification and telemicroscopy. ZooKeys 209:193–202.

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

Wilkinson M.D., Dumontier M., Aalbersberg I.J.J., Appleton G., Axton M., Baak A., Blomberg N., Boiten J.-W., Silva Santos L.B. da, Bourne P.E., Bouwman J., Brookes A.J., Clark T., Crosas M., Dillo I., Dumon O., Edmunds S., Evelo C.T., Finkers R., Gonzalez-Beltran A., Gray A.J.G., Groth P., Goble C., Grethe J.S., Heringa J., Hoen P.A.C. 't, Hooft R., Kuhn T., Kok R., Kok J.N., Lusher S.J., Martone M.E., Mons A., Packer A.L., Persson B., Rocca-Serra P., Roos M., Schaik R. van, Sansone S.-A., Schultes E., Sengstag T., Slater T., Strawn G., Swertz M.A., Thompson M., Lei J. van der, Mulligen E. van, Velterop J., Waagmeester A., Wittenburg P., Wolstencroft K.J., Zhao J., Mons B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data. 3:160018.

Wilkinson M.D., Sansone S.A., Schultes E., Doorn P., Bonino da Silva Santos L.O., Dumontier M. 2018. A design framework and exemplar metrics for FAIRness. Sci. Data 5:180118.

Wink M., Botschen F., Gosmann C., Schäfer H., Waterman G. 2010. Chemotaxonomy seen from a phylogenetic perspective and evolution of secondary metabolism. Annu. Plant Rev. 40:364–433.

Winterton S.L. 2009. Revision of the stiletto fly genus *Neodialineura* Mann (Diptera: Therevidae): an empirical example of cybertaxonomy. Zootaxa 2157:1–33.

Yilmaz P., Kottmann R., Field D., Knight R., Cole J.R., Amaral-Zettler L., Gilbert J.A., Karsch-Mizrachi I., Johnston A., Cochrane G., Vaughan R., Hunter C., Park J., Morrison N., Rocca-Serra P., Sterk P., Arumugam M., Bailey M., Baumgartner L., Birren B.W., Blaser M.J., Bonazzi V., Booth T., Bork P., Bushman F.D.,

Buttigieg P.L., Chain P.S., Charlson E., Costello E.K., Huot-Creasy H., Dawyndt P., DeSantis T., Fierer N., Fuhrman J.A., Gallery R.E., Gevers D., Gibbs R.A., San Gil I., Gonzalez A., Gordon J.I., Guralnick R., Hankeln W., Highlander S., Hugenholtz P., Jansson J., Kau A.L., Kelley S.T., Kennedy J., Knights D., Koren O., Kuczynski J., Kyrpides N., Larsen R., Lauber C.L., Legg T., Ley R.E., Lozupone C.A., Ludwig W., Lyons D., Maguire E., Methé B.A., Meyer F., Muegge B., Nakielny S., Nelson K.E., Nemergut D., Neufeld J.D., Newbold L.K., Oliver A.E., Pace N.R., Palanisamy G., Peplies J., Petrosino J., Proctor L., Pruesse E., Quast C., Raes J., Ratnasingham S., Ravel J., Relman D.A., Assunta-Sansone S., Schloss P.D., Schriml L., Sinha R., Smith M.I., Sodergren E., Spo A., Stombaugh J., Tiedje J.M., Ward D.V., Weinstock G.M., Wendel D., White O., Whiteley A., Wilke A., Wortman J.R., Yatsunenko T., Glöckner F.O. 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat. Biotechnol. 29:415–420.

Zamora J.C., and 412 coauthors. 2018. Considerations and consequences of allowing DNA sequence data as types of fungal taxa. IMA Fungus 9:167–175.

Zhang J., Kapli P., Pavlidis P., Stamatakis A. 2013. A general species delimitation method with applications to phylogenetic placements. Bioinformatics 29:2869–2876.

Zompro O. 2005. Catalogue of type material of the insect order Phasmatodea, housed in the Museum für Naturkunde der Humboldt Universität zu Berlin, Germany and in the Institut für Zoologie der Martin Luther Universität in Halle (Saale), Germany. Dtsch. Entomol. Z. 52:251–290.

Zurowietz M., Langenkämper D., Nattkemper T.W. 2019. BIIGLE2Go—a scalable image annotation system for easy deployment on cruises. OCEANS 2019-Marseille. Marseille, France: IEEE, p. 1–6.