Organisms Diversity & Evolution (2020) 20:1–8
https://doi.org/10.1007/s13127-019-00428-w

ORGANISMS DIVERSITY & EVOLUTION

FORUM PAPER

Data storage and data re-use in taxonomy—the need for improved storage and accessibility of heterogeneous data

Birgit Gemeinholzer[1] · Miguel Vences[2] · Bank Beszteri[3] · Teddy Bruy[4,5] · Janine Felden[6] · Ivaylo Kostadinov[7] · Aurélien Miralles[4,5] · Tim W. Nattkemper[8] · Christian Printzen[9] · Jasmin Renz[10] · Nataliya Rybalka[11] · Tanja Schuster[5] · Tanja Weibulat[7] · Thomas Wilke[12] · Susanne S. Renner[5]

Received: 3 June 2019 / Accepted: 14 December 2019 / Published online: 20 January 2020

Abstract

The ability to rapidly generate and share molecular, visual, and acoustic data, and to compare them with existing information, and thereby to detect and name biological entities is fundamentally changing our understanding of evolutionary relationships among organisms and is also impacting taxonomy. Harnessing taxonomic data for rapid, automated species identification by machine learning tools or DNA metabarcoding techniques has great potential but will require their review, accessible storage, comprehensive comparison, and integration with prior knowledge and information. Currently, data production, management, and sharing in taxonomic studies are not keeping pace with these needs. Indeed, a survey of recent taxonomic publications provides evidence that few species descriptions in zoology and botany incorporate DNA sequence data. The use of modern high-throughput (-omics) data is so far the exception in alpha-taxonomy, although they are easily stored in GenBank and similar databases. By contrast, for the more routinely used image data, the problem is that they are rarely made available in openly accessible repositories. Improved sharing and re-using of both types of data requires institutions that maintain long-term data storage and capacity with workable, user-friendly but highly automated pipelines. Top priority should be given to standardization and pipeline development for the easy submission and storage of machine-readable data (e.g., images, audio files, videos, tables of measurements). The taxonomic community in Germany and the German Federation for Biological Data are researching options for a higher level of automation, improved linking among data submission and storage platforms, and for making existing taxonomic information more readily accessible.

Keywords Taxonomy · Accelerated species description · Machine learning tools · Image recognition · Metabarcoding · Data repositories · German Federation for Biological Data

✉ Birgit Gemeinholzer
birgit.gemeinholzer@bot1.bio.uni-giessen.de

[1] Systematic Botany, Justus Liebig University Gießen, Heinrich-Buff Ring 38, 35392 Giessen, Germany

[2] Zoological Institute, Technische Universität Braunschweig, Mendelssohnstraße 4, 38106 Braunschweig, Germany

[3] Phycology, Faculty of Biology, University of Duisburg-Essen, Universitätsstraße 2, 45141 Essen, Germany

[4] Institut de Systématique, Evolution, Biodiversité, Muséum National d'Histoire Naturelle, Sorbonne Université, 25 rue Cuvier, 75005 Paris, France

[5] Systematic Botany and Mycology, University of Munich (LMU), Menzingerstraße 67, 80638 Munich, Germany

[6] MARUM - Center for Marine Environmental Sciences, University of Bremen, Leobener-Straße 8, 28359 Bremen, Germany

[7] GFBio - Gesellschaft für Biologische Daten e.V., c/o Research II, Campus Ring 1, 28759 Bremen, Germany

[8] Biodata Mining Group, Faculty of Technology, Bielefeld University, PO Box 100131, 33501 Bielefeld, Germany

[9] Department of Botany and Molecular Evolution, Senckenberg Research Institute and Natural History Museum Frankfurt, Senckenberganlage 25, 60325 Frankfurt/Main, Germany

[10] DZMB – Senckenberg am Meer, Martin-Luther-King Platz 3, 20146 Hamburg, Germany

[11] Experimentelle Phykologie und Algenkulturen, University Göttingen, Nikolausberger Weg 18, 37073 Göttingen, Germany

[12] Animal Ecology and Systematics, Justus Liebig University Gießen, Heinrich-Buff Ring 26, 35392 Giessen, Germany

Introduction

Due to intensive anthropogenic influences in many regions of the world, we are losing species faster than we can detect and name them (Díaz et al. 2019; Tedesco et al. 2014; Fisher et al. 2018). Because scientific (Linnaean) names are the basic units in conservation management, biosecurity, research, and legislation, only formally named species are usually receiving attention in these fields (Tedesco et al. 2014). Being able to identify and name organisms at the species level—and to do so with high speed and quality—is a pressing necessity (Wheeler et al. 2012). Taxonomists face the Herculean task of documenting, illustrating, and describing the diversity of organisms on Earth, newly discovered and known species alike.

Taxonomy is increasingly integrative, relying on many kinds of data. In contrast to other biological disciplines in which new (mostly more precise) data types often completely replace old ones, data used in taxonomy may retain their significance over centuries. Comparisons of geographical occurrence records or morphological traits, such as hair, scale, or stomata density as well as co-occurrence, like host-parasite-relationships, generate additional value for our understanding of biological, evolutionary, and ecological processes. Since the Renaissance, the sustainable maintenance of such data has been ensured by books and other print media and, of course, by physical specimens stored in natural history collections.

Computerization of science and society over the last 40 years has led to profound transformations, increasing the speed of data traffic and opening unprecedented opportunities to quickly generate and transmit visual and acoustic information, the latter highly important for the study of birds and amphibians. This transformation has also affected taxonomy, which has undergone a profound transformation, triggered mostly by molecular-genomic information that is fundamentally altering understanding of evolution and speciation. Moreover, the growing demand for taxonomic information in other branches of science has raised the need for speedy species identification (defined as the decision that an organism either belongs to an already named species or requires description and naming), which demands the combination and cross-linking of datasets or different layers of data. Automation of at least part of the taxonomic workflow, while maintaining comparability with historical data and knowledge, is needed if we are to speed up species naming and identification (of already named species). However, automated species identification by machine learning tools or DNA metabarcoding techniques can only work with comprehensive character comparison and integration with previous information. Recognition of regularities and irregularities in sets of characters provides density estimations to discover data clusters, but needs expert verification and control.

There are currently two million species names and hundreds of millions of specimens in natural history collections (Short et al. 2018), with about 20,000 new species being named each year (Zhang 2011; but see next paragraph for a finer-scale analysis). Natural history collections and libraries have started digitizing specimen catalogs and historical literature. Many name-bearing-type specimens are represented by high-resolution images on the web. Over one billion species occurrence records are available through the Global Biological Information Facility (gbif.org, accessed May 2019), and numerous curated species databases have been compiled for the "Catalogue of Life" (catalogueoflife.org; accessed May 2019). Although most genetic data are not generated by taxonomists, sequence information in the INSDC databases (www.insdc.org: GenBank, https://www.ncbi.nlm.nih.gov/genbank/), European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena), DNA Data Bank of Japan (DDBJ, http://www.ddbj.nig.ac.jp, all accessed May 2019) is growing fast and represents a valuable resource for taxonomists. In fact, sharing of molecular data, as imposed by many journals, is an excellent example of how scientific research benefits from data exchange at a global scale.

Importance of digital availability of primary data

The management (storage and retrievability) of data packages (Fig. 1) produced in alpha-taxonomic studies—dealing with taxonomic revision, the "sinking" of synonyms, and description of new species—is, however, not keeping pace with these modern developments. The number of species descriptions per year has increased for fungi and vertebrates, but has been decreasing for plants and insects for some time (Fig. 2A). Authors in the journal *Zootaxa*, here used as a benchmark, have published 13% of all animal species descriptions from 2004 onwards (Index of Organism Names, http://www.organismnames.com, accessed June 2019). Of these descriptions, many do not make reference to molecular data (Fig. 2B), and in the year 2018, only about 14,000 DNA sequences deposited in GenBank were linked to *Zootaxa* papers. With 2321 papers published in *Zootaxa* in 2018, this corresponds to an average of six DNA sequences per taxonomic study. The use of DNA information in alpha-taxonomic work differs, however, among groups of organisms, as illustrated by a survey we undertook of 2208 papers published in 2018 (Fig. 2B): While taxonomists working on fungi heavily relied on molecular data, this was only true for a small proportion of taxonomic studies on plants and, especially, insects (Fig. 2B). Information in DNA sequences is ideally suited to be used in crisp diagnoses of type material, and this is increasingly being recognized and should be promoted in all Codes of Nomenclature, not just those for fungi and prokaryotes (Renner 2016). Among the 2208 surveyed papers, genome-scale data sets (such as RADseq, Sequence capture, RNAseq,

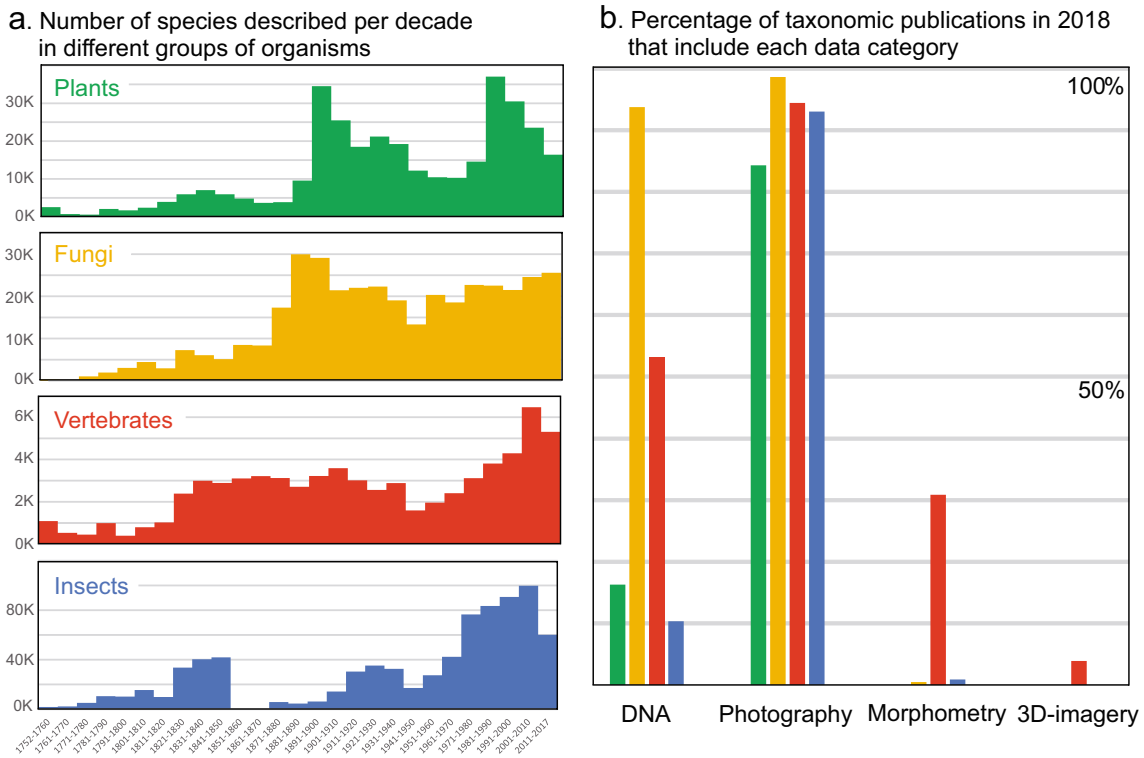

Fig. 1 (a) Species descriptions per decade in selected major groups of organisms (based on data extracted from the International Plant Names Index (https://www.ipni.org/) for plants, MycoBank (http://www.mycobank.org/) for fungi, Index of Organism Names (http://www.organismnames.com/) for insects, and compiled from various databases for vertebrates: Eschmeyer Catalog for fishes (https://www.calacademy.org/scientists/projects/eschmeyers-catalog-of-fishes), the Amphibian Species of the World for amphibians (http://research.amnh.org/vz/herpetology/amphibia/), Reptile Database for reptiles (http://www.reptile-database.org/), Howard and Moore Database (https://www.howardandmoore.org/) for birds, and ASM Mammal Diversity Database (https://mammaldiversity.org/) for mammals, all accessed in April 2019. Note that data for fungi and vertebrates refer to currently accepted species names only whereas insect data also include synonyms and subspecies, and contain data gaps for several decades in the nineteenth century; for all taxa, the low values obtained for the last decade are due to its incompleteness as data for several years are missing. (b) Histograms on the right indicate the average proportion of alpha-taxonomic studies that have produced in 2018 different categories of data. These data are based on a survey of 2208 articles (224 articles dealing with fungi, 1034 with insects, 704 with plants and 246 with vertebrates), published in 2018 in six journals with a strong taxonomic component, namely *Mycological Progress*, *Phytotaxa*, *Phytokeys*, *Systematic Botany*, *Zookeys*, and *Zootaxa*. "DNA" refers to mitochondrial or nuclear multilocus datasets, "Photography" to classical photography plus pictures generated by light and electronic 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 studies having generated data using tomographic methods (mostly 3D X-ray μCT-scan, plus one paper using synchrotron radiation μCT-scan)

full genomes) were only used in one paper in mycology (a draft genome), and not at all in zoology or in botany. Additional kinds of -omics data were also used in very minor proportions only (one publication using NIR spectra in entomology, one using NMR spectra in mycology, and one using peptide fingerprint in vertebrate zoology).

This limited use of molecular and other high-throughput data in papers that describe new species may suggest that taxonomists have not yet fully embraced the opportunities offered by DNA-sequencing, and even less of -omics approaches, or that may be working under financial and technical constraints that do not permit them to do so. Most likely, however, the reason is simply that for the largest groups of animals and plants (most of which consist of tropical organisms), there are too few DNA data to compare the already known species to potentially new ones. Therefore, most descriptions and diagnoses of new species are still based on phenotypic data alone, and this will remain unchanged unless a concerted effort is made to sequence type material of as many species as possible for a small set of standard markers. Also, at the moment, all Codes of Nomenclature require text, in addition to a specimen, for a new species to be named; the image of a DNA sequence or a link to DNA sequences in a data bank, linked to the type specimen, is insufficient for naming a new species. The storage of non-molecular raw data associated with species descriptions or taxonomic revisions also has remained problematic, and such data are therefore often lost for future re- and meta-analyses (Schmidt et al. 2016). For instance, the journals *Phytotaxa* and *Zootaxa* do not provide a default option for the publication of supplementary primary data, which could be tables, spreadsheets, images, audio files, or other kinds of data. Data packages related

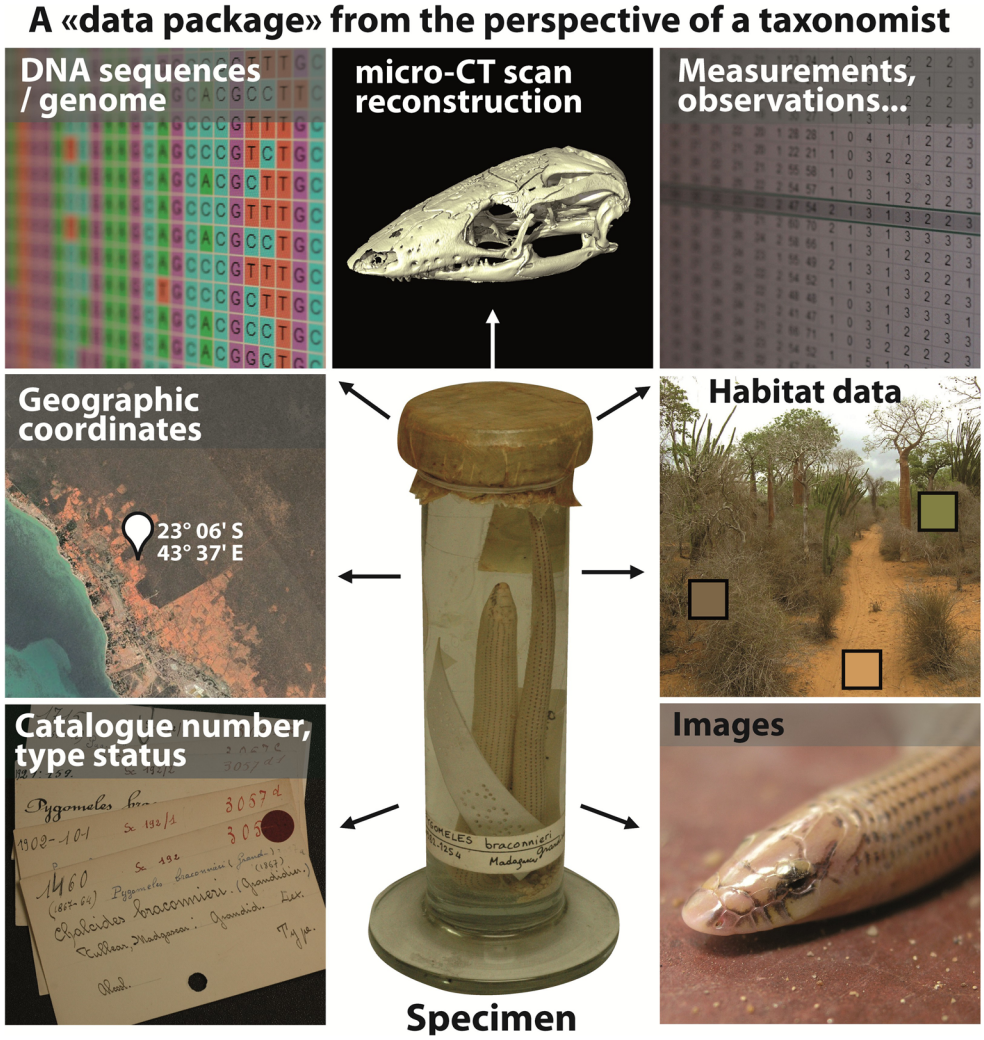

Fig. 2 Scheme of a data package in taxonomy centered around a specimen. This example is from zoology, and data types will obviously differ among taxa

to a taxonomic study can be diverse (Fig. 1) and require being stored in different distributed repositories. Other journals, such as *PhytoKeys* (http://www.phytokeys.com) and *ZooKeys* (http://www.zookeys.com), do offer this possibility and have set new standards in taxonomic publishing and especially dissemination, in full compliance with the current Codes of Nomenclature. These journals use semantic mark up, or tagging, a method that assigns markers, or tags, to text strings such as taxonomic names, gene sequences, localities, and designations of nomenclatural novelties (Penev et al. 2010, 2018). Tagged information can then be saved in machine-readable languages like XML (eXtensible Markup Language). Semantic tagging allows not only computerized methods of archiving and data mining from articles but also provides the basis for so-called semantic enhancements, which permit the linking to related articles or data. The post-publication "liberation" of text and images used in taxonomic descriptions (e.g., via the Open Biodiversity Knowledge Management System [OBKMS, Penev et al. 2010, 2018] integrated in the plazi workflow (www.plazi.org) also is a useful tool for data availability in taxonomy. While the "open access" debate is outside the scope of this paper, a key desideratum for taxonomic data is that they should also be accessible to people from developing nations and that the national science foundations of developed nations should support such accessibility financially.

The importance of raw data in taxonomy is immense, given that it is a specimen- and character-based discipline. Species, the main biological units in alpha-taxonomy, are hypotheses of how specimens—representatives of individual organisms—can best be grouped into meaningful biological and evolutionary units. The definition of "specimen" varies among groups of organisms; in zoology, it typically is an individual voucher specimen stored in a collection. In botany, it is a complete or partial plant stored in an herbarium. In paleontology, it may be a stone including one or more fossils

stored in a collection. For prokaryotes, protists and fungi, the preserved entities usually are either dried herbarium specimen or strains cultivated in a laboratory and preserved in culture collections; in other cases, microscopy slides or samples of multiple individuals might be treated as a specimen. Despite heterogeneous species concepts and problems in applying these across the Tree of Life, the availability of as many data per specimen as possible will benefit future taxonomic work. While archiving all these data may seem a massive task, we suspect it is feasible because most species descriptions do not yet include large data volumes.

Studies classifying organisms at the species level still mainly consist of text—and as revealed by our survey (Fig. 2B)—they rarely are accompanied by tables with measurements, DNA or RNA sequences, information on chemical compounds, geographical coordinates, or environmental information. One essential and widely used extra data type in taxonomy is imagery. High-resolution or low-resolution photographs of specimens can reduce the need to examine physical specimens, saving travel costs, energy and time, and such images are massively produced by taxonomists (Fig. 2B). Next-generation imagery, however, is much more powerful. In taxonomic research, robotic high-resolution 2D and 3D imaging technologies and sophisticated computer-assisted image stitching are increasingly used for the routine digitization of objects (Balke et al. 2013). Image capture can be at high resolution with 500 megapixels and more and then deliver amazing detail. The images can be used by taxonomists to store and retrieve visually accessible information about a group or a species; in addition, images can be used for computational and statistical analyses of traits.

For images to be maximally useful they should include the following information whenever possible: (a) a scale bar arranged in the images (adjacent to the sample) so that the pixel to mm ratio can be determined and (b) a color or gray value reference sheet next to the sample so that algorithms can normalize image color and contrast. This is not only necessary to align image features and assess quality for expert visual inspection, but also as a pre-processing step for computational image analysis, e.g., for automatic character and feature extraction, and (c) minimum metadata descriptions like where, when, how, and who took the image. Virtual herbaria have already implemented these recommendations extensively (e.g,. http://herbarium.bgbm.org/object/B100209261).

The storage capacity requirements for routine archiving of taxonomy-related images are high, but not unrealistically so. Assuming that 100 images were associated with each species description, about 2 million images would have to be stored per year, clearly, a manageable number even for the scientific community compared with the 50 billion pictures currently commercially hosted by Instagram alone. Moreover, generating digital image data and videos does not cost much nor require much training or expensive technologies, as indicated by the continuously rising use of citizen-science tools, such as Flora Incognita (https://floraincognita.com), iPlant (https://en.wikipedia.org/wiki/IPlant_Collaborative), and Pl@ntNet (https://plantnet.org/en/). This makes image capture an interesting tool for Citizen Science projects that could help collect digital imagery of species over large areas or at longer intervals. For posterior detailed taxonomic assignment, online tools, such as the next generation image annotation tool BIIGLE (www.biigle.de), could be used (Langenkämper et al. 2017).

Sound recordings from microphones or hydrophones for taxonomic investigations are typically converted to digital data, which demand low-storage capacity. Acoustic data storage requires a minimal set of metadata related to sensors, data acquisition, and analysis (Furnas and Callas 2015; Sevilla and Glotin 2017). Acoustic filters like Acoustic Complexity Index, Acoustic Evenness Index, Acoustic Entropy Index, and Normalized Difference Soundscape Index (Bradfer-Lawrence et al. 2019) can reduce the enormous complexity of the soundscape of recordings into biophony (noise produced by the fauna), anthrophony (human and machine noise), and geophony (noise from natural processes such as wind and rain) and allow categorizing the soundscape or identifying certain species. Currently, acoustic monitoring is time-consuming, necessitates expert knowledge and, in the case of automated recognizers, is still subject to high error rates (Furnas and Callas 2015; Sevilla and Glotin 2017). Estimations of population densities and acoustic localization relying only on acoustic recordings are still a challenge (Bradfer-Lawrence et al. 2019). Advances in standardized collection and processing methodologies however offer an enormous potential for rapid and cost-effective automated biodiversity assessments at large spatial and temporal scales, and potentially at low costs.

Benefits of data sharing in taxonomy

Taxonomists rarely work in large teams; for instance, half of the 2311 papers published in *Zootaxa* in 2018 were authored by only one or two researchers. Such small teams may lack institutional support for performing elaborate bioinformatic tasks, and submitting data to repositories may become an additional time-consuming burden. However, the benefits of data availability and the possibility of re-use by other researchers are large, both for the progress of the discipline as a whole, and for the individual researcher who aims at recognition and visibility for his or her work. For the past eight years, the National Science Foundation, USA (NSF), has sponsored the 10-year 100-million USD program "Advanced Digitization of Biodiversity Collections," which has paid for nearly 62 million specimens to be digitally photographed from multiple angles for specific research studies. It is now expanding efforts to develop a standardized, upgradable

system for linking disparate databases to create "extended specimens." Users will see not only a specimen but also data, such as DNA sequences and environmental information from its collection site. Such information can help assess shifts in faunas and floras that occurred with changes in climate or land use.

Unfortunately, despite at least a decade of calls for image sharing (e.g. Pyle et al. 2008; Seltmann 2008; Winterton 2009), the datasets underlying classical drawings or image documentation of categorical character states are rarely published. This requires the next taxonomist working on the same group of organisms to re-score the very same specimens for the very same characters, forgoing the opportunity to continuously optimize taxonomic treatments (Favret 2014). Digitally available information and comprehensive data sharing among taxonomists will not only accelerate taxonomic research but also improve the quality of the studies based on these data sets (Enke et al. 2012). Publication of data sets underlying taxonomic revisions might also benefit individual researchers by increasing both the visibility and the credit for their work. Piwowar et al. (2007) showed that studies with accompanying data sets were cited significantly more often than comparable studies where the data sets were not made public, independent of author, or journal impact factor.

To assure quality of, and access to, data, many funding agencies, such as the NSF and the DFG (Deutsche Forschungsgemeinschaft, Germany), require proposals to have data management plans describing the acquisition, processing, and storage of data collected and generated within the scope of a research projects. The NSF has required such plans since January 2011, and all such plans are reviewed as an integral part of the proposal. For a researcher to receive continued NSF funding, data availability from previous projects is checked by the reviewers. An imitation of the outstanding success of data publication by the mutualism of journals and INSDC is currently unfortunately not feasible for other data. Nevertheless, an additional incentive for digital data storage is the option of publishing data sets and their description in vetted repositories (e.g. www.pangaea.de) and in new data journals, such as *Biodiversity Data Journal* (https://bdj.pensoft.net/), *Scientific Data* (www.nature.com/sdata/), or *GigaScience* (https://academic.oup.com/gigascience), which increases the availability, visibility, and reusability of the data.

Sustainable data repositories for long-term data storage and re-use

Sharing and re-using data requires companies or institutions that maintain sustainable long-term data storage and ensure the technical implementation of data storage capacity (Bach et al. 2012; Allison et al. 2015), a task complicated by the lack of long-term funding for the relevant databases. For example, extensive efforts are required to fulfill standards in this field, such as mirroring data, as successfully implemented by the INSDC databases that store molecular data. Moreover, long-term data storage is only reasonable if the re-use of the stored data is adequately facilitated, which for alpha-taxonomy is determined by the ease of submitting data, linking data, and finding data. These factors, in turn, depend on the curation level of the archived data. Because taxonomists are often not experienced data submitters, in many cases, they also will need assistance during the submission process, which might exceed the capacity of data centers. This calls for the development of tailored, user-friendly pipelines for fast, reliable, and easy transfer of specimen-based data to the most adequate repositories.

Linking and finding species is best achieved via specimen identifiers that can be linked to characters and metadata (Güntsch et al. 2017; Groom et al. 2017; Triebel et al. 2018). Ideally, identifiers will be globally unique and persistent, unambiguously linking the respective voucher specimens and associated metadata, including a standardized protocol on how the data were collected. Several identifier systems for biodiversity research have been proposed and are in use (see lists in Guralnick et al. 2015 and Güntsch et al. 2017). The geological community has agreed to use International Geo Sample Numbers (IGSNs; http://www.geosamples.org/igsnabout), which provide not just global uniqueness but also mining authority, governance, and a set of services which is available for each registered scientist. A comparable system might work for taxonomy and biodiversity research. However, due to the infancy of globally unique specimen identifier systems, their use is not an obligatory requirement in alpha-taxonomic research. Also, because such identifiers do not yet exist for the majority of collections, it will take time before such unique identifiers will be extensively and routinely used in taxonomic publications.

In general, to convince taxonomists to archive original data, flexibility and simplicity of the submission process are crucial. Minimum requirements for taxonomic data should follow the "better than nothing" approach of the MIMARKS (minimum information on a marker gene sequence) and the MIxS standards (Minimum Information about any (x) Sequence) of the Genomic Standards Consortium (GSC; Yilmaz et al. 2011), while allowing for more detailed metadata in predefined or custom fields.

Repositories facilitate not only long-term archival, accessibility, and discovery but also can be used to monitor the quality, usage, and impact of data (Güntsch et al. 2017). Many different data repositories exist, including general-purpose ones like Dryad (https://datadryad.org) and Figshare (https://figshare.com), topic-based like MorphoBank (https://morphobank.org), journal-associated like MorphoMuseuM (https://morphomuseum.com) and GigaDB (http://gigadb.org), MorphoSource (https://www.morphosource.org), and PLAZI (http://plazi.org/resources/treatmentbank/). However,

the level of compliance of databases with the FAIR data principles, i.e., for data to be findable, accessible, interoperable, and reusable (Wilkinson et al. 2016), is not always obvious to the data submitters who may not be aware of all the facts concerning databases' eligibility, registration, copyright notice, limitations, and exceptions.

Current and future challenges and perspectives

The importance and benefits of sustainable data maintenance in taxonomy require stable and reliable repositories and workable, user-friendly, and ideally highly automated pipelines for depositing data packages. Priority should be given to standardization and pipeline development for machine-readable data, including images, audio files or videos, and tables of measurements. With increasing data availability and the development of tools for the analysis and visualization of complex biological data new, sophisticated analyses will become feasible (e.g., https://bivi.co/visualisations; http://www.paraview.org, https://vistrails.org), helping taxonomists to describe the astonishing diversity of species on Earth and to make the existing taxonomic information as accessible as possible.

As one possible option to attain this goal, the German Federation for Biological Data (Diepenbroek et al. 2014) could be made more suitable for the taxonomic community. GFBio was established as a service-oriented infrastructure network to support the data life cycle in biodiversity, ecology, and environmental science. Its data archival, publication, and discovery services are built as a single point of contact for the users of long-term data centers, including natural science collections, environmental data publishers, and the European Nuclear Archive (ENA). So far, few taxonomic data sets have been published through the GFBio services (Bruy et al. 2019; Rakotoarison et al. 2019), and taxonomist make up about 5% of the users of GFBio (own data, 12 Feb. 2019). Experience with alpha-taxonomic data packages is growing, however. The published data are linked to the GBIF portal, where their impact via citation metrics is visible. For example, the Bruy et al. (2019) dataset (https://doi.org/10.15468/l8qg0g) has been downloaded > 50 times in less than 2 months after publication. Despite such success stories, specific aspects of taxonomic data packages still need to be addressed, including the application of templates, which might be community or technology specific. GFBio and the taxonomic community in Germany are currently exploring possible solutions, which we hope will provide test cases triggering a wider adoption of data storage strategies by taxonomists.

Acknowledgements 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 advice and data on species description rates.

Funding Information Open Access funding provided by Projekt DEAL. This work benefited from the sharing of expertise within the DFG priority program SPP 1991 Taxon-Omics and support from DFG grants VE 247/16-1, RY 173/1-1, WI 1902/14-1, BE 4316/7-1, NA 731/9-1, RE 603/26-1, RE 603/29-1, GE 1242/19-1, and DFG support for GFBio DI 1219/6-3, GL 553/5-1.

Data availability The datasets analyzed during the current study are not publicly available because they only present metadata of already published data. However, data are available from the corresponding author on reasonable request.

References

Allison, L., Gurney, R., Cesar, R. M., Gemeinholzer, B., Koike, T., Mathieu, P. P., Mokrane, M., Nativi, S., Peters, D., Samors, B., Treloar, A. E., Vilotte, J-P., Visbeck, M. & Waldmann, C. (2015). A place to stand: e-Infrastructures and Data Management for Global Change Research, GEOSS Common Infrastructure: The Discovery and Access Broker (DAB) framework. Belmont Forum e-Infrastructures and Data Management Project, https://doi.org/10.5281/zenodo.34370.

Bach, K., Schäfer, D., Enke, N., Seeger, B., Gemeinholzer, B., & Bendix, J. (2012). A comparative evaluation on technical solutions of long-term data repositories in integrative biodiversity research. *Ecological Informatics, 11*, 16–24.

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., Fikáček, 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'. *Frontiers in Zoology, 10*, 55.

Bradfer-Lawrence, T., Gardner, N., Bunnefeld, L., Bunnefeld, N., Willis, S. G., & Dent, D. H. (2019). Guidelines for the use of acoustic indices in environmental research. *Methods in Ecology and Evolution, 10*(10), 1796–1807. https://doi.org/10.1111/2041-210X.13254.

Bruy, T., Vences, M., Glaw, F. & Miralles, A. (2019). A detailed morphological dataset outlining the diversity of the genus *Mimophis* (Serpentes: Psammophiinae). [Dataset]. Version: 20190325. Data Publisher: Staatliche Naturwissenschaftliche Sammlungen Bayerns – SNSB IT Center, München. http://biocase.snsb.info/wrapper/querytool/main.cgi?dsa=GFBio201900216SNSB, https://doi.org/10.15468/l8qg0g.

Díaz, S., Settele, J., Brondízio, E., et al. (2019). IPBES-7: Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, https://www.ipbes.net/event/ipbes-7-plenary.

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 E. Plödereder, L. Grunske, E. Schneider, & D. Ull (Eds.), *Informatik 2014 – Big Data Komplexität meistern. GI-Edition: Lecture Notes in Informatics (LNI) – Proceedings* (Vol. 232, pp. 1711–1724). Bonn: Köllen Verlag.

Enke, N., Thessen, A., Bach, K., Bendix, J., Seeger, B., & Gemeinholzer, B. (2012). The user's view on biodiversity data sharing. *Ecological Informatics, 11*, 25–33.

Favret, C. (2014). Cybertaxonomy to accomplish big things in aphid systematics. *Insect Sci., 21*, 392–399.

Fisher, M. A., Vinson, J. E., Gittleman, J. L., et al. (2018). The description and number of undiscovered mammal species. *Ecology and Evolution, 8*(7), 3628–3635.

Furnas, B., & Callas, R. (2015). Using automated recorders and occupancy models to monitor common forest birds across a large geographic region. *Journal of Wildlife Management, 79*, 325–337. https://doi.org/10.1002/jwmg.821.

Groom, Q., Hyam, R., & Güntsch, A. (2017). Stable identifiers for collection specimens. *Nature, 33*, 546.

Güntsch, A., Hyam, R., Hagedorn, G., Chagnoux, S., Röpert, D., Casino, A., Droege, G., Glöckler, F., Gödderz, K., Groom, Q., Hoffmann, J., Holleman, A., Kempa, M., Koivula, H., Marhold, K., Nicolson, N., Smith, V. S., & Triebel, D. (2017). Actionable, long-term stable, and semantic web compatible identifiers for access to biological collection objects. *Database*. https://doi.org/10.1093/database/bax003.

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 bio-collections data. *ZooKeys, 494*, 133–154. https://doi.org/10.3897/zookeys.494.9352.

Langenkämper, D., Zurowietz, M., Schoening, T., & Nattkemper, T. W. (2017). BIIGLE 2.0 - browsing and annotating large marine image collections. *Frontiers in Marine Science, 4*, 83. https://doi.org/10.3389/fmars.2017.00083.

Penev, L., Kress, W. J., Knapp, S., Li, D.-Z., & Renner, S. S. (2010). Fast, linked, and open – the future of taxonomic publishing for plants: launching. *The Journal PhytoKeys, 1*, 1–14. https://doi.org/10.3897/phytokeys.1.642.

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. *Biodiversity Information Science and Standards, 2*, e25748. https://doi.org/10.3897/biss.2.25748.

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS One, 2*(3), e308.

Rakotoarison, A., Scherz, M. D., Bletz, M. C., Razafindraibe, J. H., Glaw, F. & Vences, M. (2019). Media and additional measurements belonging to the description of *Cophyla fortuna (*Microhylidae, Cophylinae). [Dataset]. Version: 1.0. Data Publisher: Zoological Research Museum Koenig - Leibniz Institute for Animal Biodiversity. https://doi.org/10.20363/media-cophyla-fortuna-1.0.

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. *Systematic Biology, 65*(6), 1085–1095.

Schmidt, B., Gemeinholzer, B., & Treloar, A. (2016). Open data in global environmental research: the Belmont Forum's open data Survey. *PLoS One, 11*(1). https://doi.org/10.1371/journal.pone.0146695.

Seltmann, K. (2008). Digital image vouchering in Morphbank, linking to publications, and a few words about sharing. *American Entomologist, 54*(4), 235–238.

Sevilla, A. & Glotin, H. (2017). Audio bird classification with inception-v4 extended with time and time-frequency attention mechanisms. In CEUR Workshop Proceedings, 1866.

Short, A. E. Z., Dikow, T., & Moreau, C. S. (2018). Entomological collections in the age of big data. *Annual Review of Entomology, 63*, 513–530.

Tedesco, P. A., Bigorne, R., Bogan, A. E., et al. (2014). Estimating how many undescribed species have gone extinct. *Conservation Biology, 28*(5), 1360–1370.

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*. https://doi.org/10.1093/database/bax096.

Wheeler, Q. D., Knapp, S., Stevenson, D. W., et al. (2012). Mapping the biosphere: exploring species to understand the origin, organization and sustainability of biodiversity. *Systematics and Biodiversity, 10*, 1–20.

Wilkinson, M. D., Dumontier, M., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*. https://doi.org/10.1038/sdata.2016.18.

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., et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology, 29*, 415–420.

Zhang, Z. Q. (2011). Accelerating biodiversity descriptions and transforming taxonomic publishing: the first decade of Zootaxa. *Zootaxa, 2896*, 1–7.