Digital accessible knowledge: Mobilizing legacy data and the future of taxonomic publishing

Susan Fawcett[1], Donat Agosti[2], Selina R. Cole[3,4], David F. Wright[3,4]

1 DIGITAL ACCESSIBLE KNOWLEDGE

In the face of the modern biodiversity crisis, effectively prioritizing conservation efforts and mitigating extinction can only be accomplished with a more complete understanding of Earth's past and present biodiversity (Barnosky et al. 2011; Wilson 2017). Despite centuries of taxonomic discovery, an estimated 86 to 91% of eukaryotic species remain unknown to science (Mora et al. 2011). Taxonomic research and publications are necessary for documenting new species discoveries, updating existing species concepts, and advancing other crucial components of biodiversity knowledge, including morphology, distribution, evolutionary relationships, and keys to identification. Most commonly, taxonomic publications take the form of monographs, floras, faunas, and journal articles, but many barriers stand in the way of making the data they contain widely available. Furthermore, legacy publications contain vast amounts of biodiversity data, but this information can be difficult to access and time-consuming to extract, which places major restrictions on the feasibility of synthetic biodiversity studies. Making these data accessible increases their value (Miller et al. 2012). Here, we discuss the challenges that surround these two key aspects of biodiversity literature: the mobilization of legacy data and the future of taxonomic publishing. We provide a series of recommendations and suggested workflows to make past, present, and future taxonomic data available as digital accessible knowledge (DAK), which is defined as primary data that are both digital and accessible in standard formats (Sousa-Baena et al. 2014).

We consider the vast body of scientific literature

Affiliations:

[1]University and Jepson Herbaria, University of California, Berkeley, 1001 Valley Life Sciences Building, Berkeley, CA 94720, USA; [2]Plazi, Zinggstr. 16, 3007 Bern, Switzerland; [3]Smithsonian Institution, National Museum of Natural History, 10th St. & Constitution Ave. NW, Washington, DC 20560; [4]American Museum of Natural History, Central Park West, New York, NY 10024, USA

Correspondence:
Donat Agosti
Email: agosti@plazi.org

documentingbiodiversityknowledgetobethe universal heritage of the global community; therefore, this knowledge should be free and available to all. We advocate for taxonomic studies to apply the FAIR principles: that data—including treatments, tables, figures, bibliographic references, material citation, and methods—should be Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016). The DAK format emphasizes that data should be structured in a way that maximizes accessibility and reproducibility. It is designed to be both human- and machine-readable, including domain specific semantics that facilitate finding, citing, and linking to cited resources such as figures, earlier treatments, taxonomic keys, gene sequences or specimens. As such, DAK is the ideal format for achieving our vision of mobilizing all published taxonomic data to create a comprehensive catalogue of life.

Taxonomic monographs, the foundation of biodiversity knowledge for hundreds of years, are incredibly rich sources of data. These data include taxonomic treatments, comprehensive lists of supporting literature, figures, tables, and material citations with abundant links to external resources. As such, monographs are essentially complex, outwardly-linking citation systems. To maximize accessibility, legacy publications should be converted to DAK format so that data can be extracted for use and archived in databases, and future publications should be structured as DAK at the time of publication. This will ensure that all facts contained within monographs are easily findable and citable and that all the cited facts include their respective identifiers, thereby enriching the biodiversity citation network well beyond publications. For example, including comprehensive lists of synonymy for taxa in the form of treatment

citations is necessary, since these names are key data required for creating a catalogue of life (treatmentbank.org). By becoming an integral part of global information services, such as the Global Biodiversity Information Facility (GBIF), they will attain their intrinsic, fundamental role in the digital age.

2 DIGITAL ACCESSIBLE KNOWLEDGE: BARRIERS AND PROGRESS

Traditionally, taxonomic and associated biodiversity information has been presented as a discrete, static narrative due to the inherent constraints of print media. Although the transition to online publishing has been slow, the opportunities presented by online formats are revolutionizing the practice of taxonomy (Godfray et al. 2007; Kress and Penev 2011; Marhold et al. 2013; Côtez et al. 2018). The raw data fundamental to taxonomy are rapidly becoming accessible through coincident digitization efforts led by natural history museums all over the world. These resources (with representative examples in Table 1.) include: digitized legacy taxonomic literature; data extracted and made FAIR from articles, especially taxonomic treatments; semantically enhanced publications; digitized natural history collection data, including type specimens; observational data, including photographs from the field; and genomic sequence data. Many of these resources, now representing more than 1.9 billion occurrence records, are aggregated in the Global Biodiversity Information Facility (GBIF 2021), facilitating the development of the extended or digital specimen, sensu Webster (2017) and Hardisty et al. (2019). For example, a specimen may be linked, using its persistent identifier, to duplicates at other

institutions, gene sequence data, taxonomic treatments and other literature that cite it. The aggregation and linking of these data present opportunities for new avenues of research (Heberling et al. 2021) while also revealing gaps in existing knowledge (Lughandha et al. 2019; Marshall et al. 2018) and enabling the re-use of data in synthetic ways (Clark et al. 2009; Heberling et al. 2019; Heberling et al. 2021).

Taxonomic publications are enhanced by unlimited access to data and literature (Orr et al. 2021). A major impediment to accessing and digitizing printed taxonomic literature is copyright law. Copyright law may be unique and applicable to variable time periods from nation to nation but generally places a restriction on the reproduction, sharing, copying, or distribution of publications for several to many decades after publication. However, as argued by Agosti and Egloff (2009), copyright law applies to "literary and artistic" work but does not apply to data or "facts" that can be shared openly. Access can also be assured by obtaining individual licenses from the publishers or authors (e.g., BHL) or, if possible, signing contracts with collective societies that will reimburse the authors (e.g., tinyurl.com/tam5r9jz). All the data may then be made open and FAIR,

Name	Resources Available
Biodiversity Heritage Library (BHL)	digitized legacy taxonomic literature
Hathi Trust	digitized legacy taxonomic literature
Internet Archive	digitized legacy taxonomic literature
TreatmentBank	FAIR taxonomic data from publications
Biodiversity Literature Repository (BLR)	FAIR taxonomic treatments and figures from publications
Biodiversity Community Integrated Knowledge Library (BiCIKL)	FAIR linked biodiversity data
Pensoft Publishers	semantically enhanced publications
European Journal of Taxonomy	semantically enhanced publication
Integrated Digitized Biocollections (iDigBio)	digitized natural history collection data
National Specimen Information Infrastructure (NSII)	digitized natural history collection data
National Research Collections Australia (NRCA)	digitized natural history collection data
JSTOR	digitized type specimens and literature
iNaturalist	biodiversity observation data
BioScan	genomic sequence data
Earth BioGenome Project (EBP)	genomic sequence data
European Reference Genome Atlas (ERGA)	genomic sequence data
National Center for Biotechnology Information (NCBI)	genomic sequence data
Sequence Read Archive (SRA)	genomic sequence data
Global Biodiversity Information Facility (GBIF)	biodiversity data aggregator
Atlas of Living Australia (ALA)	biodiversity data aggregator
Paleobiology Database (PBDB)	fossil biodiversity database

Table 1. Examples of taxonomic and biodiversity data resources.

including the respective license. However, the best way to avoid future problems is by publishing open access.

Converting taxonomic publications to digital accessible knowledge can be achieved most efficiently by using semantically enhanced publishing workflows (Kress and Penev 2011). If this is not a feasible near-term solution, a service to convert traditional monographs (in PDF-format) into DAK can be used. Providing clear formatting guidelines to authors—like those provided by the European Journal of Taxonomy (Chester et al. 2019) and adopted by Pensoft journals—will greatly facilitate conversion to DAK. Fortunately, monographs generally are semantically highly structured and predictable (Miller et al. 2012), which makes them ideal for conversion and data enhancement (Fig. 1).

3 MOBILIZING LEGACY DATA: DISCOVERING KNOWN BIODIVERSITY

In recent years, there has been a substantial increase in both archiving and using data from online biodiversity data repositories within the biological sciences (Edwards et al. 2000; Heberling et al. 2021). These repositories host many types of biodiversity data, the majority of which are extracted from monographs and other taxonomic publications. While a wide range of tools, platforms, and workflows have been developed to facilitate this work, widespread use of these resources in standardized ways has not been adopted (Bayraktarov et al. 2019). Further, these databases are often highly incomplete, and the time-consuming nature of extracting data from taxonomic publications remains a major barrier to synthetic biodiversity studies (Kissling et al. 2015). As a result, a major objective for

legacy taxonomic data should be identifying tools to extract and mobilize unstructured published data to make it easily accessible for researchers. We argue that mobilizing legacy data is a key step towards the ultimate goal of creating a comprehensive catalogue of all life, including synonyms, that is (1) linked to all cited scientific data, (2) hosted on one or more centralized, sustainable platforms with links that connect and synchronize with associated platforms, where appropriate and, (3) fully accessible following FAIR data guidelines.

The benefits of such a goal are numerous. Extraction of data from publications allows paywalls to be avoided, facilitating universal access to taxonomic and other biodiversity data from anywhere by anyone at any time. This would increase accessibility of taxonomic data to both professional researchers and avocational scientists around the globe. A centralized platform and/or use of common formats and vocabularies to share and provide access to decentralized storage allows data reuse, aids synthetic studies, and accelerates research. While many platforms for storing and accessing biodiversity data currently exist, many of these are taxon-specific, have limited access, or are difficult to integrate and maintain (e.g., Moudrý and Devillers 2020). Finally, once these data have been extracted and made freely available, they offer extensive benefits to taxonomists and researchers across the biological sciences. These up-to-date resources are of critical value to land managers, conservationists, policy makers, and other stakeholders.

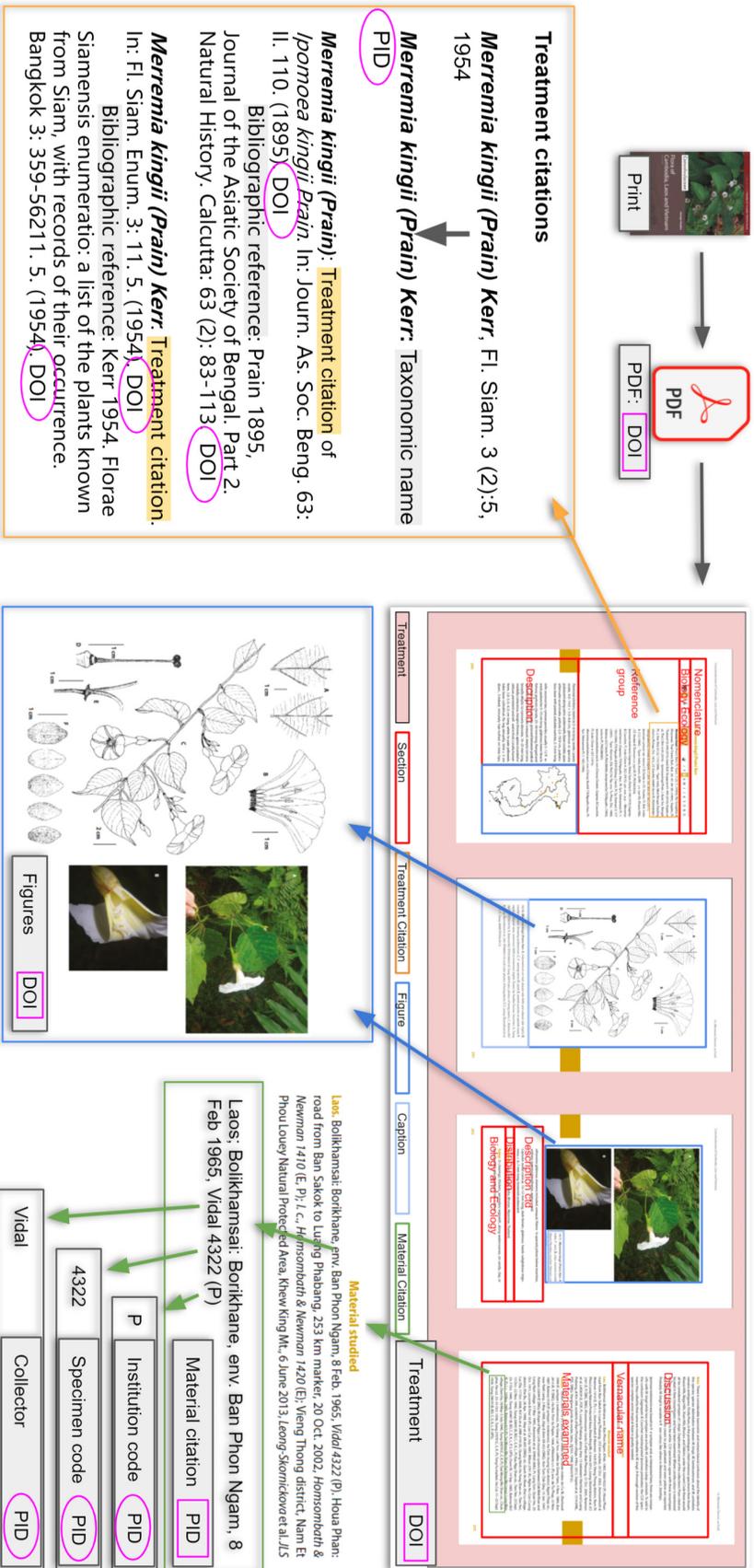## 4 EXAMPLE WORKFLOW FOR CONVERTING TAXONOMIC LITERATURE TO DIGITAL ACCESSIBLE KNOWLEDGE

Fig. 1. The wealth of digital, accessible, citable knowledge that is hidden in a single taxonomic treatment and imprisoned in a printed flora. This example is the treatment of Meremia kingii (Prain) Kerr published in the print only volume of the Convolvulaceae in the Flora of Cambodia (Staples 2018). (source: sciencepress.mnhn.fr/en/thematics/flora-cambodia-laos-vietnam). DOI: Digital Object Identifier; PID: Persistent Identifier.

In order to contribute to a global biodiversity knowledge graph sensu Page (2016) and to support future monography, publication data must be made open and FAIR. This requires not only for the data to be discovered, enhanced, and stored in a local database, but also for it to be uploaded to respective infrastructures and assigned persistent identifiers using universal vocabularies for the metadata. Such a service is provided by Plazi, a Swiss not-for-profit association dedicated to supporting and promoting the development of persistent and openly accessible digital taxonomic literature (plazi.org). Plazi developed and maintains TreatmentBank (TreatmentBank 2009), a workflow and service to convert and extract data from scholarly publications (Fig. 2). Plazi and Pensoft co-founded the Biodiversity Literature Repository community (Biodiversity Literature Repository 2013) at Zenodo, to provide long term access to these extracted FAIR data (Agosti and Egloff 2009).

The input can be anything, from a hard copy to scanned publications to XML or born digital Portable Document Format (PDF) publications. These documents are then converted into a text stream, that includes figures or multimedia content with captions linked to figure citations in the text to allow extractions of text, without losing the connection to the figure. The next step is to extract the article metadata, with or without retrieving and comparing it to the metadata obtained from the CrossRef DOI resolution service. This is followed by enhancement of the bibliographic references by linking them to their sources as well as to within-text citations. Taxonomic names are identified, normalized, and annotated with the vocabulary and hierarchy obtained from the taxonomic backbone at GBIF and the Catalogue of Life (https://

catalogueoflife.org/). The sections of the article are semantically enhanced with an additional step for further subdividing treatments to recognize elements such as nomenclature, descriptions, material examined, or conservation assessments. As an additional step, material citations (i.e., citation of specimens examined) are tagged and their content used to annotate them so they can be linked to and made citable from specimens, gene sequences, collectors, or institutions. Treatment citations are tagged and normalized as a basis for building the catalogue of life, and, if possible, linked to the cited treatment. Each of these tags is assigned a unique identifier (UUID). Collection, specimen, and accession codes are, if possible, identified, and, if available, the persistent identifier of the code is attributed to the respective annotation. A quality control tool helps to filter the data and identify any necessary corrections. The data will then be released to users, such as GBIF or BLR, based on predefined criteria that correspond to their specific needs.

The result will be stored as a file in the non-proprietary Image Markup File (IMF) format, which is similar to the star-schema used in Darwin Core Archive. For each page, it includes a reference image used for the coordinate system to define the position of each token (word). A system of CSV files then includes the structural and semantic information for the entire document based on the individual tokens. Multimedia files of each figure and graphic are included as well as the original PDF file. Upon upload of the file to the TreatmentBank server, the data are imported into a database (Postgres), and the article, figures, and treatments are deposited to BLR, which generates a DOI for each deposit that will be added to the respective annotations (e.g., the DOI for a figure links to the figure caption and figure citations).
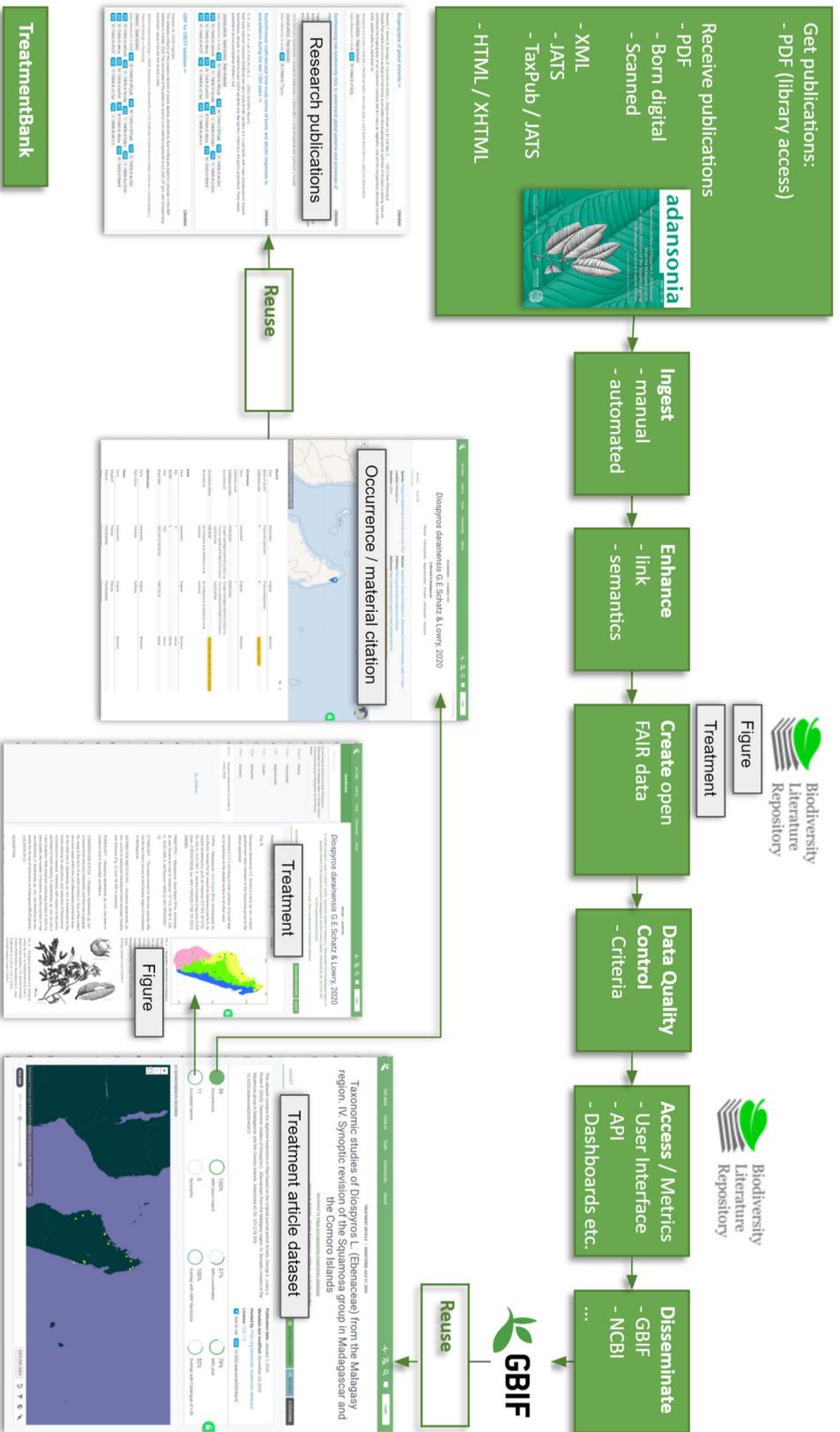
Fig 2. TreatmentBank workflow to convert unstructured taxonomic research data into digital accessible knowledge. Source: Article: Schatz and Lowry 2020; BLR:zenodo.org/record/3953000; GBIF: www.gbif.org/dataset/4f2bbc27-03f2-46a2-a461-9995a8a5a5fd; GBIF reuse: www.gbif.org/resource/search?contentType=literature&gbifDatasetKey=4f2bbc27-03f2-46a2-a461-9995a8a5a5fd

Once this is completed and the quality control shows that the data are fit for use, a DarwinCore Archive is created including only the individual treatments and material citation, which are imported by GBIF. After successful upload to GBIF, the respective GBIF identifier for the article deposit will be embedded in the metadata of the article, as well as in the metadata of the BLR deposit. For closed access articles, only the data are open access; the article itself is not accessible, but the metadata of its deposit will be.

The entire workflow is based on widely used data vocabularies in the biodiversity community (e.g., Darwin Core, TDWG) or Taxpub JATS (Journal Article Tag Suite), which has been specifically developed for publishing taxonomic data. This allows third parties to develop tools to import data into GBIF, or to adopt it for new publications.

Plazi is collaborating with the European Journal of Taxonomy to develop publishing guidelines (Chester et al. 2019) to ease conversion of taxonomic publications to DAK. Many of these guidelines have now been adopted by Pensoft publishers (e.g. PhytoKeys c2020).

This entire workflow does not and will never operate entirely error-free without human intervention, and its products will not be fit for each user. For that reason, feedback mechanisms are in place. GBIF users send messages from within the platform or contact Plazi via its community issue tracker. This feedback will be used to fix errors and, at the same time, will help to improve the processing by adjusting the underlying algorithms.

From this point of view, it is also clear that the best strategy for the future is to structure monographic publications so as to avoid the need for processing. This is exemplified by Pensoft publishers' 25 journals, which are available as DAK in BLR and GBIF the moment they are published.

5 RECOMMENDATIONS FOR FUTURE TAXONOMIC PUBLISHING

We envision that future taxonomic publications will be intrinsically linked to all supporting data and literature. We recognize taxonomic classifications to be scientific hypotheses and, as such, the datasets supporting them should be reproducible. This is possible when all examined specimens, molecular vouchers, cited literature, and supporting datasets are digitized and linked within the document using persistent identifiers or DOIs (digital object identifiers); in other words, they are digital accessible knowledge. Monographs can become living documents with dynamic distribution maps that can be replaced by updated versions (and previous versions archived) as more data become available. Alternatively, they can be a starting point that can be augmented with additional publications, which are ideally linked bidirectionally.

In many ways, our recommendations to increase data accessibility in taxonomic research integrates the practice of monography into the broader trend toward "open science" policies in biology. Although description is the heart of taxonomy, there are many forms of associated data included in monographic publications falling outside the realm of pure description. At a minimum, we advocate monographs be published in machine-readable formats to capture and preserve this information using a framework like the one discussed above. However, we can envision ways taxonomists can take a page from our colleagues in related, data-driven disciplines, such as computational biology and ecology. In these fields, it is

increasingly commonplace to ensure all data and code are publicly available (Hampton et al. 2015; Parker et al. 2016). Often, free online repositories are used to store this information (e.g., GitHub), which raises the possibility of creating "living documents" of data, methods, and code while enhancing reproducibility using version control. Where possible, we encourage taxonomists to take similar steps to make species data, specimen metadata, and all associated information (e.g., trait measurements, geographic occurrence data, etc.) available in free online repositories (e.g., Dryad, Zenodo). We believe these efforts would complement, not supplant, the practice of taxonomy by creating a more open community of scientists and enhance data recovery and reproducibility (Wilkinson et al. 2016).

We advocate for the new Bulletin of the Society of Systematic Biologists to publish taxonomic data and associated information in the form of digital accessible knowledge (DAK). It is clear this can not be done in one step, so we recommend the following:

1. Publish open access.
2. Provide clear guidelines and templates to authors for publishing in a structured format that will allow their data to be quickly and easily extracted and included by data aggregators (see guidelines in Penev et al. 2012; Penev et al. 2017; Chester et al. 2019).
3. Facilitate the creation of XML documents by providing user-friendly article submission portals that include categorical components (e.g., taxonomic treatment, synonyms, description, diagnosis, key, material citation [with spreadsheet template], ecology, conservation assessment, miscellaneous notes, etc.).
4. Cite all bibliographic references in full, or

include a DOI so that citation networks can be built.
5. Use existing persistent identifiers when available for specimens, species, gene sequences, taxonomic treatments, figures, tables, phylogenies, and publications (e.g., Güntsch et al. 2017; Klump and Huber 2017; McMurry et al. 2017; Juty et al. 2020).
6. Generate persistent identifiers for those elements that do not yet have them.
7. Maintain data structure by ensuring all data tables and associated information are published in a machine-readable format (e.g., Vogt 2019).
8. Archive FAIR data on an open, accessible online repository so that it can exist as a companion resource to the associated publication(s) and as a "living document."

6 CONCLUSIONS

Taxonomic literature, especially monographs, provides the foundation for identifying known biodiversity as well as a framework for the discovery and description of unknown biodiversity (Grace et al. 2021). Historically, both natural history collections and taxonomic literature have been largely inaccessible to the general population. Making this knowledge accessible through digitization will allow for a larger and more diverse community of taxonomists, especially from countries rich in biodiversity and from populations that have been excluded historically (Drew et al. 2017). By empowering this broader community and facilitating discovery of biodiversity, taxonomic literature in the form of digital accessible knowledge is an indispensable tool for combating biodiversity loss.

Acknowledgements

Daly (The Ohio State University), and all participants of the NSF-sponsored workshop on "Collaborative Research: Revolutionizing Systematics - Revitalizing Monographs" DEB-1839205. We thank Bruce Baldwin (JEPS), Torsten Dikow (NMNH), and an anonymous reviewer for helpful comments on the manuscript.

References

Agosti D, Egloff W. Taxonomic information exchange and copyright: the Plazi approach. BMC Res Notes. 2009;2:53. https://doi.org/10.1186/1756-0500-2-53

Barnosky AD, Matzke N, Tomiya S, Wogan GOU, Swartz B, Quental TB, Marshall C, McGuire JL, Lindsey EL, Maguire KC, Mersey B, Ferrer EA. Has the Earth's sixth mass extinction already arrived?. Nature. 2011;471(7336):51–57. https://doi.org/10.1038/nature09678

Bayraktarov E, Ehmke G, O'Connor J, Burns EL, Nguyen HA, McRae L, Possingham HP, Lindenmayer DB. Do big unstructured biodiversity data mean more knowledge?. Front Ecol and Evol 2019;6(239), 1–5. https://doi.org/10.3389/fevo.2018.00239

Biodiversity Literature Repository. Zenodo. https://zenodo.org/communities/biosyslit/?page=1&size=20. c2013 [cited 2021 May 04].

Chester C, Agosti D, Sautter G, Catapano T, Martens K, Gérard I, Bénichou L. EJT editorial standard for the semantic enhancement of specimen data in taxonomy literature. Eur J Taxon. 2019;(586): 1–22. https://doi.org/10.5852/ejt.2019.586

Clark BR, Godfray HCJ, Kitching IJ, Mayo SJ, Scoble MJ. Taxonomy as an eScience. Philos Trans A Math Phys Eng Sci. 2008;367(1890):953–966. https://doi.org/10.1098/rsta.2008.0190

Côtez E, Mabille A, Chester C, Rocklin E, Deroin T, Desutter-Grandcolas L, Lesur J, Merle D, Robillard T, Bénichou L. 1802–2018: 220 ans d'histoire des périodiques au Muséum. Adansonia. 2018;40(1):1–40. https://doi.org/10.5252/adansonia2018v40a1

Drew JA, Moreau CS, Stiassny ML. Digitization of museum collections holds the potential to enhance researcher diversity. Nat Ecol Evol. 2017;1(12):1789–1790. https://doi.org/10.1038/s41559-017-0401-6

Edwards JL, Lane MA, Nielsen ES. Interoperability of biodiversity databases: biodiversity information on every desktop. Science. 2000;289(5488):2312–2314. https://doi.org/10.1126/science.289.5488.2312

GBIF. New data-clustering feature aims to improve data quality and reveal cross-dataset connections. https://www.gbif.org/news/4U1dz8LygQvqIywiRIRpAU/new-data-clustering-feature-aims-to-improve-data-quality-and-reveal-cross-dataset-connections. c2020 [cited 2021 Apr 29].

Godfray HCJ, Clark BR, Kitching IJ, Mayo SJ, Scoble MJ. The web and the structure of taxonomy. Syst Biol. 2007;56(6):943–955. https://doi.org/10.1080/10635150701777521

Grace OM, Pérez-Escobar OA, Lucas EJ, Vorontsova MS, Lewis GP, Walker BE, Lohmann LG, Knapp S, Wilkie P, Sarkinen T, Darbyshire I, Lughadha EN, Monro A, Woudstra Y, Demissew S, Muasya AM, Díaz S, Baker WJ, Antonelli A. Botanical Monography in the Anthropocene. Trends Plant Sci. 2021;26(5):433–441. https://doi.org/10.1016/j.tplants.2020.12.018

Güntsch A, Hyam R, Hagedorn G, Chagnoux S, Röpert D, Casino A, Droege G, Glöckler F, Gödderz K, Groom Q, Hoffmann J. Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. Database. 2017;2017(bax003):1–9. https://doi.org/10.1093/database/bax003

Hampton SE, Anderson SS, Bagby SC, Gries C, Han X, Hart EM, Jones MB, Lenhardt WC, MacDonald A, Michener WK, Mudge J, Pourmokhtarian A, Schildhauer MP, Woo KH, Zimmerman N. The Tao of open science for ecology. Ecosphere. 2015;6(7):1–13. https://doi.org/10.1890/ES14-00402.1

Hardisty, AR, Ma K, Nelson G, Fortes J. (2019) 'openDS'–A new standard for digital specimens and other natural science digital object types. Biodiversity Information Science and Standards. 2019;3:e37033. https://doi.org/10.3897/biss.3.37033

Heberling JM, Prather LA, Tonsor SJ. The changing uses of herbarium data in an era of global change: An overview using automated content analysis.

BioScience. 2019;69(10):812–822. https://doi.org/10.1093/biosci/biz094

Heberling JM, Miller JT, Noesgaard D, Weingart SB, Schigel D. (2021) Data integration enables global biodiversity synthesis. Proc Natl Acad Sci U S A. 2021;118(6):e2018093118. https://doi.org/10.1073/pnas.2018093118

Juty N, Wimalaratne SM, Soiland-Reyes S, Kunze J, Goble CA, Clark T. Unique, persistent, resolvable: Identifiers as the foundation of FAIR. Data Intell. 2020;2(1-2):30–39. https://doi.org/10.1162/dint_a_00025

Kissling WD, Hardisty A, García EA, Santamaria M, De Leo F, Pesole G, Freyhof J, Manset D, Wissel S, Konijn J, Los W. Towards global interoperability for supporting biodiversity research on essential biodiversity variables (EBVs). Biodiversity. 2015;16(2-3):99–107. https://doi.org/10.1080/14888386.2015.1068709

Klump J, Huber R. 20 Years of persistent identifiers–Which systems are here to stay? Data Sci J. 2017;16(9):1–7. https://doi.org/10.5334/dsj-2017-009

Kress WJ, Penev L. Innovative electronic publication in plant systematics: PhytoKeys and the changes to the "Botanical Code" accepted at the XVIII International Botanical Congress in Melbourne. PhytoKeys. 2011;(6):1–4. https://doi.org/10.3897/phytokeys.6.2063

Lughadha EMN, Graziele Staggemeier V, Vasconcelos TNC, Walker BE, Canteiro C, Lucas EJ. Harnessing the potential of integrated systematics for conservation of taxonomically complex, megadiverse plant groups. Conserv Biol. 2019;33(3), 511–522. https://doi.org/10.1111/cobi.13289

Marhold K, Stuessy T, Agababian M, Agosti D, Alford MH, Crespo A, Crisci JV, Dorr LJ, Ferencova Z, Frodin D, Geltman DV, Kilian N, Linder HP, Lohmann LG, Oberprieler C, Penev L, Smith GF, Thomas W, Tulig M, Turland N, Zhang XC. The future of botanical monography: Report from an international workshop, 12–16 March 2012, Smolenice, Slovak Republic. Taxon. 2013;62(1):4–20. https://doi.org/10.1002/tax.621003

Marshall CR, Finnegan S, Clites EC, Holroyd PA, Bonuso N, Cortez C, Davis E, Dietl GP, Druckenmiller PS, Eng RC, Garcia C. Quantifying the dark data in museum fossil collections as palaeontology undergoes a second digital revolution. Biol lett. 2018;14(9):20180431. https://doi.org/10.1098/rsbl.2018.0431

McMurry JA, Juty N, Blomberg N, Burdett T, Conlin T, Conte N, Courtot M, Deck J, Dumontier M, Fellows DK, et al. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLoS Biol. 2017;15(6):p.e2001414. https://doi.org/10.1371/journal.pbio.2001414

Miller J, Dikow T, Agosti D, Sautter G, Catapano T, Penev L, Zhang Z, Pentcheff D, Pyle R, Blum S, et al. From taxonomic literature to cybertaxonomic content. BMC Biol. 2012;10:87. https://doi.org/10.1186/1741-7007-10-87

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B. How many species are there on Earth and in the ocean?. PLoS Biol. 2011;9(8):e1001127. https://doi.org/10.1371/journal.pbio.1001127

Moudrý V, Devillers R. Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data. Ecol Inform. 2020;56:101051. https://doi.org/10.1016/j.ecoinf.2020.101051

Orr MC, Ferrari RR, Hughes AC, Chen J, Ascher JS, Yan YH, Williams PH, Zhou X, Bai M, Rudoy A, et al. Taxonomy must engage with new technologies and evolve to face future challenges. Nat Ecol Evol. 2021;5(1):3–4. https://doi.org/10.1038/s41559-020-01360-5

Page R. Towards a biodiversity knowledge graph. Res Ideas Outcomes. 2016;2:e8767. https://doi.org/10.3897/rio.2.e8767

Parker TH, Forstmeier W, Koricheva J, Fidler F, Hadfield JD, Chee YE, Kelly CD, Gurevitch J, Nakagawa S. Transparency in ecology and evolution: real problems, real solutions. Trends Ecol Evol. 2016;31(9):711–719. https://doi.org/10.1016/j.tree.2016.07.002

Penev L, Catapano T, Agosti D, Georgiev T, Sautter G, Stoev P. Implementation of TaxPub, an NLM DTD extension for domain-specific markup in taxonomy, from the experience of a biodiversity publisher. In: Journal Article Tag Suite Conference (JATS-Con)

Proceedings 2012 [Internet], Bethesda (MD), National Center for Biotechnology Information (US). 2012. https://doi.org/10.5281/zenodo.804247

Penev L, Mietchen D, Chavan VS, Hagedorn G, Smith VS, Shotton D, Tuama ÉÓ, Senderov V, Georgiev T, Stoev P, et al. Strategies and guidelines for scholarly publishing of biodiversity data. Res Ideas Outcomes. 2017;3:e12431. https://doi.org/10.3897/rio.3.e12431

Author guidelines. PhytoKeys. 2020 [cited 2021 Apr 29]. https://phytokeys.pensoft.net/about#Author-Guidelines.

Schatz GE, Lowry II PP. Taxonomic studies of *Diospyros* L. (Ebenaceae) from the Malagasy region. IV. Synoptic revision of the Squamosa group in Madagascar and the Comoro Islands. Adansonia. 2020;42(10):201-218. https://doi.org/10.5252/adansonia2020v42a10

Sousa-Baena MS, Garcia LC, Peterson AT. Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. Divers Distrib. 2013;20(4):369–381. https://doi.org/10.1111/ddi.12136

Staples GW. Convolvulaceae. In: Flora of Cambodia, Laos, Vietnam, volume 36. Muséum national d'Histoire naturelle, Paris, Marseille, Edinburgh. 2018. https://doi.org/10.5852/fft47

TreatmentBank. Plazi. c2009 [cited 2021 May 04]. http://plazi.org/resources/treatmentbank/.

Vogt L. Organizing phenotypic data—a semantic data model for anatomy. J Biomed Semantics. 2019;10(1):1–14. https://doi.org/10.1186/s13326-019-0204-6

Webster MS, editor. The Extended Specimen: Emerging Frontiers in Collections-based Ornithological Research. Boca Raton, FL: CRC Press, Taylor & Francis Group; 2017. https://doi.org/10.1201/9781315120454

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3(1):160018. https://doi.org/10.1038/sdata.2016.18

Wilson EO. Biodiversity research requires more boots on the ground. Nat Ecol Evol. 2017;1(11):1590–1591. https://doi.org/10.1038/s41559-017-0360-y