European Journal of Taxonomy 1004: 120–143
https://doi.org/10.5852/ejt.2025.1004.2971

europeanjournaloftaxonomy.eu
ISSN 2118-9773
2025 · Buschbom J. *et al.*

Received: 10 March 2024 • *Accepted: 29 April 2025* • *Published: 21 July 2025*

Topic editor: Tony Robillard • *Section editor: Frank Zachos* • *Desk editor: Chris Le Coquet-Le Roux*

Research article

Open data in publications – non-copyrightability and attribution as drivers for equity, science and innovation

Jutta BUSCHBOM[1,*], Laurence BÉNICHOU[2], Donat AGOSTI[3], Willi EGLOFF[4], Elisa HERRMANN[5], Mariko KAGEYAMA[6], Andreas KROH[7] & Patricia MERGEN[8]

[1] Statistical Genetics, Ahrensburg, Germany.
[2] Muséum national d'Histoire naturelle, Paris, France.
[3, 4] Plazi, Berne, Switzerland.
[5] Museum für Naturkunde Berlin (MfN) – Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany.
[6] Research Management Center, Tohoku University, Sendai, Japan.
[7] Natural History Museum Vienna, Austria.
[8] Meise Botanic Garden, Belgium.
[8] Royal Museum for Central Africa, Belgium.

[*] Corresponding author: jutta.buschbom@statistical-genetics.de
[2] Email: laurence.benichou@mnhn.fr
[3] Email: agosti@plazi.org
[4] Email: egloff_bader@bluewin.ch
[5] Email: elisa.herrmann@mfn.berlin
[6] Email: mariko.kageyama@tohoku.ac.jp
[7] Email: andreas.kroh@nhm.at
[8] Email: patricia.mergen@africamuseum.be

Buschbom J., Bénichou L., Agosti D., Egloff W., Herrmann E., Kageyama M., Kroh A. & Mergen P. 2025. Open data in publications – non-copyrightability and attribution as drivers for equity, science and innovation. *European Journal of Taxonomy* 1004: 120–143. https://doi.org/10.5852/ejt.2025.1004.2971

Abstract. The mobilization of the wealth of existing biodiversity data is fundamental for the development of large, dynamic and multifaceted datasets and their historical baselines. Biodiversity data need to be reliably and persistently linked to their sources in literary works and in databases. For this purpose, biodiversity data in scholarly publications and their associated repositories have to be transformed to provide findable, accessible, interoperable and reusable resources as a core foundation of interlinked, federated information. Unclear rights and obligations form a substantial obstacle to the reuse and effective interlinking of data, and thus to scientific workflows. Therefore, it is key to understand and implement the legal, ethical and social foundations that are crucial for arriving at informed decisions on the access to and the transformation and reuse of such data.

The aim of this article is to provide legal clarity to providers and users of such data, and to give recommendations arising from the analysis of the legal background and of community norms requiring attribution, transparency, and accountability. The goal of the resulting recommendations is to empower the biodiversity sciences and data community, including publishers, authors and users, to apply appropriate legal tools as well as language that will provide legal certainty, thereby accelerating the access to and annotation, extraction and reuse of data contained within publications, both legacy and prospective. The paper is the outcome of a workshop organized during the annual meeting of TDWG, the Biodiversity Information Standards organization, held in Sofia, Bulgaria in October 2022. Its focus was on legal and contractual obligations governing data within works and databases. It also addressed community norms, focusing on attribution as well as ethical and sociocultural principles.

Keywords. FAIR, linked biodiversity data, public domain mark, scientific best practices and code of conduct, data governance labels.

Introduction

A key question for basic biodiversity research and conservation is how to mobilize the wealth of existing biodiversity data for the development of large, dynamic and multifaceted datasets and their historical baselines. Towards this end, the EU-funded BiCIKL project (2021–2024) focused on the implementation of bidirectional interlinking of data present both within prospective and legacy publications (Penev *et al.* 2022). The aim was to develop basic functionality for interlinking publications with data in key research infrastructures, such as the Global Biodiversity Information Facility (GBIF), TreatmentBank, the Biodiversity Literature Repository (BLR), Zenodo, the Swiss Institute of Bioinformatics Library Systems (SIBiLS), or the European Nucleotide Archive (ENA), and these infrastructures back to the publications and the data therein. Progress in this area contributes to a vision for open knowledge management (Bouchout Declaration, Anonymous 2014; Group on Earth Observations 2021; UNESCO 2021). It requires primary biodiversity data to be reliably and persistently linked to their sources in literature. For this purpose, biodiversity data and their associated repositories need to be transformed into findable, accessible, interoperable and reusable (FAIR, Wilkinson *et al.* 2016) resources as a core foundation for the generation of interlinked, federated information that facilitates transdisciplinary investigations, biodiversity management operations and fact-based policy-making.

Access to FAIR data needs to be based on well-developed systems for data governance to find acceptance by data providers and users, as well as societies and governments (e.g., RDA-CODATA Legal Interoperability Interest Group 2016; Carroll *et al.* 2020; Oldham *et al.* 2023; GO FAIR 2024). Exploring and clarifying the legal, ethical and sociocultural contexts of FAIR data in environments dominated by copyright law, Bénichou *et al.* (2023) developed a set of best practices that will foster the extraction and reuse of high quality and information-rich biodiversity data from copyrighted works, specifically scholarly publications. At the same time, the developed best practices promote the attribution of FAIR data to providers and improve transparency and accountability for their users.

Currently, most small publishers, specifically institutional and/or learned society journals in the natural sciences sector, express concerns related to copyright and are uncertain if they are allowed to share data contained within a published paper without a clear statement from the author. Similarly, many authors are also unaware of whether or not they retain copyright for their text and data in publications. Finally, legal uncertainty and cumbersome, even unmanageable, procedures widely persist for biodiversity scientists and data managers who are interested in, and dependent on, the reuse of data published in scholarly publications and digital infrastructures. Unclear rights and obligations form a substantial obstacle to the effective interlinking of data, and thus scientists' and data managers' work. Our aim is to provide legal clarity to providers and users.

This paper is the outcome of a workshop organized during the annual conference of TDWG, the Biodiversity Information Standards organization, held in Sofia, Bulgaria in October 2022. The workshop was jointly organized by members of the Biodiversity Heritage Library (BHL), the Consortium of European Taxonomic Facilities (CETAF) e-Publishing working group and the Society for the Preservation of Natural History Collections (SPNHC). The event was supported by the Biodiversity Community Integrated Knowledge Library project (BiCIKL, Penev *et al.* 2022). The focus of the workshop was on legal and contractual obligations governing data within copyright-protected works. Such data can be embedded within the main body of the publication itself, e.g., in the form of tables, figures or verbalized descriptions, or attached as supplementary data. During drafting of the Best Practices (Bénichou *et al.* 2023), the need to further contextualize and clarify the underlying rationales became apparent.

The joint CETAF, SPNHC and BHL Best Practices build on existing frameworks, as for example the Bouchout Declaration on Open Biodiversity Knowledge Management (Anonymous 2014), the Legal Interoperability of Research Data: Principles and Implementation Guidelines (RDA-CODATA Legal Interoperability Interest Group 2016), the GEO Statement on Open Knowledge (Group on Earth Observations 2021), the Recommendation on Open Science (UNESCO 2021), the Recommendation of the Council on Enhancing Access to and Sharing of Data (OECD 2021) and the CARE principles (Collective benefit, Authority to control, Responsibility, Ethics, Carroll *et al.* 2020). It considers discussions of copyright-associated questions in scientific contexts (e.g., Watanabe 2018; European Commission Directorate-General for Research and Innovation & Angelopoulos 2022). Bénichou *et al.* (2023) reinforces existing best practice guidelines in use by the biodiversity sciences and informatics community (Ball 2014; Patterson *et al.* 2014; Egloff *et al.* 2016, 2017; Bénichou *et al.* 2018, 2021) and adapts them to evolving stakeholder perspectives (Hahnel *et al.* 2023), as well as the changing legal landscape and global policy contexts of the digital transformation.

The first section of the paper outlines and discusses the different dimensions and practical questions associated with copyright in the context of the difference between the European and U.S. American legal systems, and future developments that were presented during the workshop. In a second section, we develop reliable, and legally sound approaches that can be suitable to provide legal certainty for text- and data-mining applications. On the one hand, we describe ways forward for the extraction and reuse of scientific data contained in legacy publications. On the other hand, we develop recommendations for prospective publications. By promoting explicit data governance statements by authors and publishers this paper aims at providing a foundation, based on which publishers, authors and data users following these guidelines can be confident to publish and reuse data fearless of copyright issues.

Differentiating copyright laws, scientific attribution and data governance

It is key to determine and differentiate the legal, ethical and social foundations that govern data for understanding the conditions under which data can be accessed, retrieved, reused and shared.

Copyright as a legal construct

Legal systems in the traditionally analog and the now increasingly digital publishing sectors protect the products of individual creativity by defining a bundle of rights with respect to these specific creative products. The rightsholder can control certain defined (re)uses (see for example the *Berne Convention for the Protection of Literary and Artistic Works* (WIPO 1979), Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market (European Parliament and Council 2019a) and section 106 of Title 17 of the United States Code on copyright law (Office of the Law Revision Counsel of the House of Representatives 2023a). This bundle includes the rights to reproduce, distribute, make available and communicate to the public, to modify and to generate derivatives.

However, from a legal perspective, data are not categorized as individual products of creativity, and therefore cannot be copyrighted (Gervais 2017: 94 ff). For example, the U.S. Supreme Court (1991) in a ruling holds that raw data are not copyrightable (see Feist Publications, Inc. v. Rural Telephone Service Co., 499 U.S. 340, U.S. Supreme Court 1991). In this ruling the court states that:

> "(a) Article I, § 8, cl. 8, of the Constitution mandates originality as a prerequisite for copyright protection. The constitutional requirement necessitates independent creation plus a modicum of creativity. Since facts do not owe their origin to an act of authorship, they are not original and, thus, are not copyrightable." (U.S. Supreme Court 1991: 340).

Transferred into the scientific context, this means that, from a copyright point of view, scientific data themselves are non-creative. They are characterized as qualitative or quantitative statements (Sequoiah-Grayson & Floridi 2022; Adriaans 2023) that in the sciences are the results of standardized protocols, though often with randomized elements. These routines are conducted by hand or as the output of analog and/or digital machines automatically performing actions according to a prescribed rules table. These rule-based outcomes are considered to be empirical facts, which aim to represent, that is "copy" (Paul & Stokes 2023), a pre-existing reality as neutral, accurate and unembellished as possible (Sequoiah-Grayson & Floridi 2022). Hence, the reality-prescribed data themselves are not considered to be expressions of individual ingenuity and creativity (Paul & Stokes 2023), and as such are not copyrightable. Their form and information content is dictated by reality, applicable (technical) standards and scientific best practice (Adriaans 2023). Even if published in a textual context or form, the data themselves are not the product of creative choices or expressions made by the author(s) of the publication.

The structure and text conventions of scholarly publications themselves follow prescribed best practices, in which widely agreed-on templates are followed and creativity is limited. In scientific publications, creativity and ingenuity are generally limited to the choice and design of hypotheses and their tests (cf. the ability to implement scientific processes of 'strong inference', Platt 1964). The selection and development of hypotheses and their tests, as well as the expertise to differentiate elements of observations, recognize their importance as well as their impacts for scientific conclusions and further studies can be ingenious and even creative by producing surprise and showing originality, spontaneity, and/or agency (Paul & Stokes 2023). Yet, nevertheless, this ingenuity does not affect the status of the data themselves within a publication as being not copyrightable. This standpoint is supported by U.S. law, which considers mere factual, descriptive and data-driven work barely copyrightable and provides at most uncertain ("thin") protection for such works.

In sum, no person, organization or legal entity can hold any exclusive rights over data based on copyright law. Hence, no person or organization is given the legal power to exclude others from accessing and reusing data within existing copyright legal frameworks. As an analogy, data can be deemed to correspond to words in a publication. In a publication, individual words cannot be copyrighted, while the text that is built using them is a product of individual human creation and creativity and, thus, can be copyrighted.

Legal status of data compilations and databases

While legal systems do not consider data to be copyrightable, they may treat comprehensive datasets that are original by their arrangement or form of presentation, and that are held in tangible media, e.g., databases and data infrastructures, as copyrightable.

In U.S. case law language such datasets or data assemblages are called "compilations" and U.S. case law states that:

> "A compilation is not copyrightable *per se*, but is copyrightable only if its facts have been 'selected, coordinated, or arranged *in such a way* that the resulting work as a whole constitutes an original work of authorship.' § 101 (emphasis added). Thus, the statute envisions that some ways of selecting, coordinating, and arranging data are not sufficiently original to trigger copyright protection. Even a compilation that is copyrightable receives only limited protection, as copyright does not extend to the facts contained in the compilation." (citation and emphasis as in the original, Feist Publications, Inc. v. Rural Telephone Service Co., 499 U.S. 340, U.S. Supreme Court 1991: 340).

Similarly, in the EU "original" databases are protected by copyright (Directive (EU) 96/9/EC, European Parliament and Council 1996; Huemer 2021). However, Article 3 of Directive (EU) 96/9/EC explicitly states that "The copyright protection of databases provided for by this Directive shall not extend to their contents…".

In addition to "original" databases, the EU also recognizes "non original databases" (Huemer 2021) and provides for a *sui generis* protection of them "for the maker of the database which shows that there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification, or presentation of the contents…" (Article 7 Directive (EU) 96/9/EC, European Parliament and Council 1996). Again, the *sui generis* protection aims at protecting only the database as a whole and not the data units stored within it. This is specifically the case if data extraction is "for the purposes of illustration for teaching or scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved" (Article 9 Directive (EU) 96/9/EC).

The U.S. copyright protection for data compilations as well as the EU *sui generis* protection for databases are deemed to be only thin (Reichman & Okediji 2012: 1418). Such an assessment is similar to the evaluation by Ball (2014) of the practicalities of database protection in the EU. Nevertheless, Reichman & Okediji (2012) uncovered that the EU *sui generis* protection inadvertently has the effect of resulting in a protection regime that can include the data themselves and can be more restrictive than copyright. They argued that this does not improve legal clarity and has deleterious consequences for today's international research. Already, Reichman & Uhlir (2003) discussed different approaches to database protection and concluded that consequences of any database protection would not be desirable for scientific research and innovation. Recent legislation in the EU associated with the *European Data Strategy* (COM(2020)66, European Commission 2020) has the goal to increase legal certainty and improve the availability of data for reuse (see e.g., the "Data Governance Act" (Regulation (EU) 2022/868, European Parliament and Council 2022) and the "Data Act" (Regulation (EU) 2023/2854, European Parliament and Council 2023)).

Shifting the focus to goals

To enhance legal certainty for the reuse of data from publications, it is useful to point out that an extensive and diverse gray area exists between data, which are not copyrightable, and potentially protected datasets. To cut through the filigree details, ambiguities and uncertainties of this gray area and arrive at a practical, generally applicable and usable recommendation that we think provides legal certainty, we suggest to shift the focus from the "what" (What is copyrightable?) to a pragmatic "why" (Why copyright and license a scientific outcome or product?), that is, to purposes and intended goals. This shift needs to be discussed by the community as a whole to develop a way forward, even if it only can be answered by the individuals and institutions (including publishers) that are the (potential) copyright owners of a specific creation and work.

Presumably, the reason for scientists to claim copyright and attach licenses to publications and their intent to do the same for data, in general, is to enforce recognition of researchers' contributions to hard-

gained data. Indeed, current research assessment methods rely heavily on publication-based metrics such as citation counts, and often fail to recognize the wide array of contributions made by researchers. In addition, as scientists, the priority of our actions and endeavors is the responsible reuse of data with the aim of expanding and growing knowledge and insight, as well as for applications that foster society, innovation and the finding of solutions to the challenges of today's chronic polycrisis. Thereby, attribution supports quality assessments within scientific processes and strengthens research integrity, while reuse of data further improves the findability of data. The resulting connections and increased visibility promote our contributions in a measurable manner to these efforts and solutions to society and funders (Hahnel *et al.* 2023).

Attribution as scientific best practice and community norm

A primary concern of data providers, who are often also scientists and authors of academic papers, is that their contributions are properly recognized by receiving reliable and persistent attribution as well as public visibility. Full attribution of sources and origins is best scientific practice that provides transparency and accountability (e.g., Anonymous 2014; Bénichou *et al.* 2022). In this way, attribution enables exchanges about and scrutiny of data, results and proposed scientific hypotheses and, hence, ensures scientific and research integrity (Office of Science and Technology Policy 2022). In this process of attribution, data, publications and scientists are linked to each other via references and citations, thereby giving rise to a social network that is also a knowledge infrastructure. Furthermore, scientists' own as well as science-sociological, -economic and -political assessments highly value the reuse of publications and data as indicators or proxies of scientific quality and impact (Council of the European Union 2022; Hahnel *et al.* 2023). Attribution forms the basis of the most widely applied bonus system within the scientific community for assessing a scientist's contribution to and impact on scientific progress.

Hence, apart from short-term embargo periods (which are deprecated by the U.S. government, see Office of Science and Technology Policy 2022, and within the EU, see Council of the European Union 2023), the long-term core interest of data providers in scientific contexts is not to exclude others from using data by claiming copyright. On the contrary, the open sharing of data, information and knowledge as a prerequisite for their reuse is most highly valued (European Parliament and Council 2019b; UNESCO 2021; National Science Foundation 2023). Taken together, attribution of data providers in the scientific environment ensures recognition of their efforts and expertise in making available well-researched, quality-assured densely interlinked, open and free FAIR data for their frequent, continuous and widespread reuse.

Strengthening the core value of sharing data and results through detailed attribution of sources and origins of data, information and knowledge, which are well-curated and interlinked, simultaneously contributes to and assures quality data and results. Such an approach and work practice fosters a cultural shift (see Nikander *et al.* 2020; European Commission 2021; Coalition for Advancing Research Assessment 2024) from a focus on the quantity of an undiscerned, rumbled data heap to the quality of well-curated, -structured and -maintained data and workflows made available through tools and work environments that provide valuable public services.

The attribution tools developed by the partners of the BiCIKL project (Penev *et al.* 2022) and an accompanying recommendation for fully citing taxonomic authorities in a standardized way by BHL, CETAF and SPNHC (Bénichou *et al.* 2022) contribute towards this goal. They promote and provide a basis for persistently linking data with their provenance history by fostering the implementation and use of machine-actionable attribution standards (cf. the Citation File Format and CITATION.cff approach supported by Github, Zenodo and Zotero, Smith *et al.* 2016). Through the support of automation, an opportunity opens up for the efficient and effective building and use of integrated, transparent and information-rich knowledge graphs assuring full attribution to the information sources.

Parallels potentially might be drawn between the concerns of scientists revolving around sharing, attribution and reuse of data and those of Indigenous Peoples and Local Communities (IPLC) regarding traditional knowledge: IPLCs generally are not the legal rights holders to copyright, e.g., of publications describing, referring to and using a nation's, tribe's or community's traditional knowledge. From a copyright perspective, they cannot control the (re)use of their traditional knowledge. Yet, while they may not be recognized as rights holders under the conventional copyright law theory, based on ethical principles and sociocultural norms, they can be considered as the cultural authority for traditional knowledge associated with their communities and histories. They require social solutions and systems for the implementation of sovereignty and good governance (Hudson *et al.* 2023), including, for example, inclusion, self-determination, fairness and equity, participatory decision-making, recognition and trust. Such a system has been laid out in the CARE principles, recognizing and describing sociocultural norms regarding collective benefits, the authority to control, responsibility and ethics (Carroll *et al.* 2020, 2021). The CARE principles have been made actionable by engaged members of IPLC communities and adopted by an increasing number of cultural heritage organizations, e.g., through the development of the Local Contexts labels and notices (https://localcontexts.org/) that can be associated with traditional knowledge and biocultural collections and data.

Maybe most important for the context here is a general interest of many IPLC communities in and their openness to sharing their knowledge with others and in this way contributing to the building of bidirectional connections to coming generations, neighbors, travelers and thus societies close by and far away. Historically intuitive open science approaches by IPLCs, their scientists and knowledge holders predate and at the same time shape the horizon for western societies' scientific communities' ethical and social norms. They have in common with the communities of today's Western scientists the shared cherishing of the human and intrinsic value that arises from the spread and reuse of data and knowledge. Both contexts recognize a dependence on attribution as one factor that enables the flow of subsequent benefits back to their origins.

The digital commons and its digital public goods

As a scientific community we advocate the strengthening of a digital public commons for scientific data, which are to be considered a digital public good that shall be fairly shared, equitably accessible and of use to all (UNGA 2020), not only to a few experts with extensive resources and specialist tools. Within the natural sciences community, access to data and research outcomes at its core should no longer be a matter of exclusivity and power provided by legal rights, but rather become an expression of best practices and community norms. On the foundation of data governance as a social solution there is the growing recognition that reality and the data that represent it belong to everyone, that they are a public good, and that the community should accept the responsibilities that arise from a digital public commons.

An accessible and widely usable digital public commons can be achieved and expedited through densely interlinked data and infrastructures. Digital tools as outlined in the Biodiversity Knowledge Hub (https://bicikl-project.eu/biodiversity-knowledge-hub) enable and call for the automation of attribution, immediately integrating the recording and attaching of attribution within scientific workflows. In this way, such tools will embed attribution into wider contexts of resources and their agents, and open these contexts themselves up to further inquiries and research. For example, in-depth investigations can provide insights into the contributions of women, youth and minorities, specifically from the Global South or IPLCs, see Targets 22 and 23 of the Kunming-Montreal Global Biodiversity Framework (KM-GBF, CBD 2022). A powerful digital public commons and the automatization tools that mobilize its data as its public good will in addition be essential for capacity-building, technology transfer, and scientific and technical cooperation (Target 20) as well as knowledge management and services in support of biodiversity action (Target 21) of the KM-GBF (CBD 2022).

Public goods are characterized by non-exclusivity and non-rivalry (Musgrave & Musgrave 1989). As described above, empirical facts exist independently of any human activity, including observation. Biodiversity data are products of planetary and evolutionary history and thus are part of reality. They cannot be owned or kept exclusive (non-exclusivity). As digital public goods, reality and empirical data can be accessed, shared and used by everybody, either directly or in the form of derivatives. Furthermore, a pivotal characteristic of data is that access does not reduce their availability nor can the data be deleted or consumed through access to them (non-rivalry or non-competitiveness).

Digital public goods grow with frequent and widespread access and reuse (Reichman & Uhlir 2004; Nikander *et al.* 2020; Guidi 2023; however, see also Purtova & van Maanen 2024). The scientific value and intrinsic relevance is positively correlated with the openness of the data, their frequency of use, the ability to share it widely, and to improve its quality continuously through community-based curation, annotation, interlinking and extending of the scientific data. Examples are the Global Biodiversity Information Facility (https://www.gbif.org/), the World Register of Marine Species (https://marinespecies.org/), or the International Nucleotide Sequence Database Collaboration (https://www.insdc.org/) to name just three of many community-driven biodiversity data initiatives.

Although scientific data, as digital public goods, tend to be free for all, they are still governed by best practices and norms of the scientific community. The classification as a digital public good points to a modus of sharing that avoids legal enclosure, fosters fair and equitable access and enables everyone to build upon them (Dulong de Rosnay & Stalder 2020). At the same time, these data are managed by the scientific community, which sets the rules of access and of use for the data that it holds in common (Cosens *et al.* 2021; Hudson *et al.* 2023).

As highlighted above, the general aim of governance systems for digital public goods, is not to avoid access or use. Restrictions at least in the long-term will be very costly and/or impossible to establish. This is especially the case for data, information and knowledge that are to date already available and have no ethical, social or legal concerns associated with their availability, access, reuse and wide distribution. On the contrary, the implementation of governance systems should aim to provide legal clarity that increases the reuse of data as common public goods. Providing adequate governance systems for data that carry aspects of a sensitive nature with them will broaden and improve the availability of material and digital public goods that are freely accessible and reusable, even if associated with certain conditions.

Should one decide to limit access to certain empirical data and to the distribution of knowledge about reality by attaching conditions, the implementation of such a decision would require specific, well-founded and -formed arguments that are convincing to a large majority of the scientific community so that they become best practice and community norms. Nevertheless, and of fundamental importance, initial development of such governance systems, their implementation, long-term maintenance and management require well-designed, extensive and long-term concerted and supported efforts, funding and socioeconomic resources. One concrete step towards suitable agreed and adaptive governance systems are data management plans, which currently become prerequisites as required by the authors' codes of conduct, scientific publishers, or funders.

In sum, governance systems are mechanisms and processes that support and make possible the most extensive access and reuse of existing and future data, and thus their application for societal solutions to today's challenges. A combination of providing clarity for scientific data recorded as part of copyrighted publications and adopting good data governance approaches will improve the reuse of biodiversity data and thus deploy their value as digital public goods that will contribute to new research findings as well as to processes that respond efficiently and effectively to humanity's current crises.

Use case: liberating data from legacy publications

In the following sections, the conditions for liberating data from scholarly publications, that is accessing, annotating, extracting, and reusing data, are discussed under legal aspects. In the first part, guidelines are given for the access to and reuse of data from already existing legacy publications. A second set of guidelines provides orientation to authors and publishers on how to set data governance prerequisites so that data for future, prospective publications can be accessed with increased legal certainty about their reusability, potential legal limitations or obligations, and clarity about providers' cultural contexts, needs and preferences.

Learning from the past: existing and future legislation might impede access and reusability

Our legal analysis (see the section on *Copyright as a legal construct*) leads us to the conclusion that data, in particular scientific data, are not copyrightable, even if they are part of a copyrightable publication. The way data are presented is dictated by theoretical standards, technical capacity and scientific good practice. Hence, data in themselves are not the result of creative choices made by the authors or their original expression. Accordingly, the copyright protection of a publication refers to the work, not to the data contained in it. Likewise the European *sui generis* protection for databases refers to the database as an infrastructure, not to the data units it contains. Parts of a work are only protected as far as they are works in themselves.

As a consequence, the data within those publications are open and can be freely extracted, once legal access has been gained to existing copyrighted (scholarly) publications. Furthermore, following extraction, these "liberated" data continue to remain open. Hence they can be deposited in FAIR and freely accessible repositories that provide continuous open and free access for further sharing, distribution and reuse of the data.

Liberating data from existing publications therefore means – from a copyright point of view – extracting unprotected data from protected works (or databases), often referred to as text and data mining (TDM). Definitions for TDM cannot be found for all jurisdictions, e.g., no definition exists under U.S. copyright law, and no international consensus definition seems to have been reached so far. In the context of our work, including this article and the recommendation published by Bénichou *et al.* (2023), we understand TDM as "any automated analytical technique aimed at analyzing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations", as it is defined in Art. 2 n. 2 Directive (EU) 2019/790 (European Parliament and Council 2019a).

TDM as an automated procedure includes the reuse of the protected work as a whole, as do some manual approaches that are based on local copies of the full content of a copyrighted work. Accordingly, access to and reuse of the work needs authorization. This authorization can be given by contractual license, e.g., between the publisher and a user, or by legal license. Legal licenses can be compulsory (i.e., they are applicable even where the parties concerned have stipulated otherwise) or subsidiary (i.e., they are only applicable as far as the parties have not stipulated otherwise).

Actual legislation differs from country to country, so that copyright legislation today presents an international patchwork that includes legal divides:

Within the European Union directive (EU) 2019/790 (European Parliament and Council 2019a) has introduced two legal licenses referring to text and data mining: Art. 3 obliges every member state to introduce into its national copyright law a compulsory legal license enabling text and data mining for the purposes of scientific research conducted by recognized research organizations and cultural heritage institutions. Art. 4 obliges them to introduce a subsidiary legal license for any other form of text and data mining for any other purpose. Hence, in the EU, extracting data from publications for the purposes of

scientific research is allowed by law for a research institution "to conduct scientific research or to carry out educational activities involving also the conduct of scientific research: (a) on a not-for-profit basis or by reinvesting all the profits in its scientific research…" (Directive (EU) 2019/790 Art. 2 n. 1). This authorization prevails over any contractual agreement and also over any attached licenses (including e.g., CC-licenses).

The French ordonnance of Nov. 24, 2021 (Ministère de la Culture 2021), transposing the European directive, goes one step further and allows content protected by intellectual property rights, including copyright, to be reproduced for the purposes of scientific research by anybody. Therefore, in France, not only scientists at nationally accredited research institutions, but also private citizens, businesses and publishers are able to reuse a work in whole for TDM with a scientific objective, once legal access to the work has been gained. It is not necessary in France to obtain prior authorization or licenses for scientific TDM analyses, including (automated) data identification and extraction, from rights holders. It is in this exemption from authorization that the TDM "exception" defined by French law lies.

In Switzerland, extracting data from publications is allowed by legal license since a revision of Swiss copyright law in 1992 (SR 231.1, Bundesversammlung der Schweizerischen Eidgenossenschaft 1992). Due to this TDM-friendly national law, the not-for-profit organization Plazi (http://plazi.org/) based its extraction workflow in Switzerland. Systematic extraction of taxonomic data from scientific publications started in 2009. From 2013 and onwards, the extracted data are deposited in the Biodiversity Literature Repository (https://biolitrepo.org/) in Zenodo. Zenodo (https://zenodo.org/) is a FAIR and open general-purpose repository developed under the European OpenAIRE program (https://www.openaire.eu/) and operated by CERN (Conseil européen pour la recherche nucléaire, https://home.cern/). Hence, data are deposited in an open repository that at least implicitly follows the TRUST principles for digital repositories (Transparency, Responsibility, User focus, Sustainability, Technology, Lin *et al.* 2020). The results are data, which are freely and FAIRly accessible and reusable from there on, and hence available with a long-term perspective. There has never been any dispute referring to an alleged copyright infringement.

In the United States, the same TDM procedures may require a prior license to gain lawful access to and make a copy of a copyrighted work from which data are to be extracted. Such a prior license (e.g., in the form of a use agreement) is needed, unless those who are going to text-mine the copyrighted material can objectively prove that their use is "fair use" in defense against a plausible infringement claim alleged by the work's copyright holder. The question remains unresolved as to whether the act of making temporary copies of protected works merely as factual datasets for TDM projects counts as the act of reproduction within the U.S. copyright law context. Whereas the EU's approach of a categorical safe harbor aims at striking a defined balance between the protection of copyright holders and the promotion of scientific innovation driven by TDM, the U.S.'s fair use doctrine seems sufficiently flexible in its application despite a higher perceived risk of infringement. The U.S. law relies on rather fact-specific case-by-case analysis as to whether a particular use represents fair use or not. U.S. case law employs a four-factor test, which is applicable to all types of protected works to limit the exclusive rights statutorily granted to copyright holders pursuant to 17 U.S.C. § 107 (Office of the Law Revision Counsel of the House of Representatives 2023b): (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work.

Under all jurisdictions, the main principle remains the same: once extracted in adherence with local national legislation, the data can be reused freely. From a copyright point of view, the use of these extracted data worldwide does not need further authorization. Legal restrictions may apply only as far as

they are arising from other protection schemes (see e.g., UNESCO 2021) such as those concerning, for example, the protection of national security, the right to privacy, or the protection of endangered species.

At this point, with the data existing beyond any confines of copyright law, we would like to reiterate that sociocultural and ethical behavior should not be confused with copyright-imposed conditions or liberation from copyright. Scientific best practices, community norms and ethical principles exist independently and remain untouched by the copyright status of data's contexts. Hence, it is scientific best practice to attribute the extracted data to their source of extraction and original provider(s). This reciprocal practice mutually benefits the scientific community and societies' aims to enable, foster and support science.

Deficiencies of the new regulations for text and data mining

Existing legal contexts and the user roles and purposes that they define (i.e., acting as research institution or not, in a (non-)commercial legal context, for the purpose of profit or not) can introduce an additional layer of uncertainty and constraints to data reuse. Furthermore, they might introduce and reinforce existing and potential inequalities and biases towards certain actors and sectors. One example is EU directive (EU) 2019/790 of the European Parliament and Council (2019a) on copyright in the digital single market, containing provisions related to TDM discussed in the previous section. It distinguishes data reuse by (non-commercial) research organizations for the purpose of not-for-profit (non-commercial) scientific research from other types of reuse and categories of users.

While the situation is mostly clear for scientists permanently employed at accredited research institutions, the situation provides no legal security for a wide range of actors in the diversified research sectors contributing to and relying on biodiversity and environmental data. For many of these actors, it is not clear if from a legal point of view they are considered to be acting in the role of a researcher in a suitably recognized research context as required by the EU directive 2019/790 and its transpositions into national law. The question arises, for example, for employees of non-governmental organizations, free-lancing independent scientists and professionals, as well as many small to larger businesses existing in the biodiversity domain. These businesses include, e.g., consulting companies that are performing environmental, ecological and biodiversity assessments and thereby contributing indispensable experience, data, information and knowledge to governmental planning and decision processes.

Furthermore, specifically for business members of the research community, it is difficult to identify if the main purpose of their work, including the reuse of data, is primarily scientific or instead technical (cf. Information-Communication-Technology consulting), or if it needs to be considered commercial use. Here, the commercial aspect might be limited to earning a livelihood and running a small consulting business. The directive remains vague on criteria that unambiguously categorize the roles of researchers, the types of scientific work and the scientific purpose(s), which would enable actors to reliably assess if they fall within the scope of the compulsory license of Art. 3 of Directive (EU) 2019/790 that provides a safe harbor for scientific TDM use.

Stepping beyond specific national legal contexts and approaching data reuse in general as well as TDM specifically, from a worldwide, multinational research perspective, deciding which national legislation is applicable for which data set or subset within a publication, also known as choice of law or conflict of laws questions, can add further complexity. Consider, for example, an international, non-EU citizen postdoc who is employed elsewhere in the world and comes to the EU on a fellowship to work in an EU Member State (MS) on a research project for a specific period of time. While working at the accredited host research institution, the postdoc accesses and extracts data from publications within the context of the EU Member State. They carry out TDM-associated work not only while on-site at the research project's host institution based in the EU Member State. Furthermore, they are not only

using information and communication technologies (ICT)-infrastructure exclusively located in the EU. They also take advantage of the functionalities and power of globally distributed cloud servers and/or carry out TDM-associated work in part or total while working remotely abroad, outside the EU, e.g., during field research, stays in their home country or during travels to third countries. The principle question is whether EU law is applicable throughout their work on the research project. Which legal factors should be considered to determine the local national context in an era of parallel part-time employments and fellowships, remote work and institutions depending on cloud server infrastructures and capacities?

Finally, individual research projects are often part of larger, ongoing collaborations, in which serial funding phases form large-scale research initiatives of longer duration. Will a (foreign) scientist, whether on a short-term contract at a research institution located in an EU Member State or tenured at an accredited research institution in a country outside Europe and working as fellow or guest researcher at an EU-based research organization, be required to delete different stages of raw full-text and data-containing elements extracted from copyrighted publications once short-term funding is over and/or they leave the EU, respectively? EU directive 2019/790 Articles 3(2) and 4(2), as well as transposing national laws (e.g., see German copyright law UrhG §60d (4) and (5), Deutscher Bundestag 2021) remain vague in light of today's complex, multistep and globally international research workflows and practices, and thus don't seem to provide sufficient legal certainty. For example, automated TDM may require several optimization cycles for machine extractions and associated analyses that continue beyond the original project phase, e.g., to provide provenance information and, hence, transparency in support of reproducibility. Restrictions on the long-term retention of copies (see, e.g., Directive (EU) 2019/790 and subsequent national transposing laws of EU Member States) might force a scientist to (1) reproduce time-consuming steps within a workflow; (2) stop providing access to raw data resources to project partners who subsequently develop into informal collaborators; (3) disentangle data at different stages of cleaning and machine-actionability or with different modifications for analytical optimizations from interlinked data in infrastructure systems; and (4) find ways to delete the data from local hard drives, institutional networks and backup systems.

A pragmatic way forward in the currently ambiguous and patchy legal landscape in which published works exist will be to voluntarily set copyrightable works into the public domain so that they can be openly and freely accessed and reused.

Recommended best practice for future prospective publications

The described lack of legal clarity and security deters stakeholders from implementing much needed industry-strength, large-scale TDM initiatives (Plazi 2023). These initiatives could efficiently detect, extract and transfer data from copyrighted works into open repositories, achieving FAIR status for the extracted high-quality biodiversity data. Retaining the effective linkage between the extracted biodiversity data and their publications of origin, the outcomes of TDM workflows would form reliable nodes and links for the growth and expansion of urgently needed knowledge graphs. Such workflows are already successfully employed by a growing number of projects, for example, Bionomia (https://bionomia.net/) and bITEM (https://bitem.at/).

To accelerate and upscale TDM workflow development and achieve extensive knowledge graphs by overcoming existing limitations and obstacles for mobilizing high-quality biodiversity data, we recently published a set of four recommendations aimed at authors and publishers of future, to-be-published works (Bénichou *et al.* 2023). These recommendations have the support of the scientific and professional communities represented by BHL, CETAF and SPNHC that are closely involved in and depend on TDM and the mobilization of biodiversity data.

Moreover, the recommendations align with and are in the spirit of *A European strategy for data* (COM(2020)66, European Commission 2020) and derived legislation. For example, Directive (EU) 2019/1024 (PSI and Open Data Directive, European Parliament and Council 2019b) establishes that publicly funded research data should be "open by default" and if concerns exist, research data should be "as open as possible, as closed as necessary" (Art. 10). Other legal documents of interest in this context are, for example, Regulation (EU) 2022/868 (Data Governance Act, European Parliament and Council 2022) and Regulation (EU) 2023/2854 (Data Act, European Parliament and Council 2023).

The four recommendations published by Bénichou *et al.* (2023) are:

(1) authors and publishers make copyrighted publications as accessible as possible by waiving copyright (CC0) or publishing with a CC-BY license;

(2) authors and publishers explicitly state that they consider scientific data as not copyrightable (see the detailed Blue List republished by Bénichou *et al.* 2023). Best practice is to attach to the data a public domain mark that provides certainty about their open and free reusability;

(3) publishers use a publishing technique that is supporting automatic text and data mining (Agosti *et al.* 2022);

(4) authors state as clearly and comprehensively as possible the provenance of their data, the authors of previous works cited (Bénichou *et al.* 2022), and – for works having more than one author – the respective contributions of all co-authors, e.g., using the CRediT system (https://credit.niso.org/).

The primary aim of these four recommendations for future works as outcomes of scientific research is to provide legal clarity and certainty to all involved stakeholders associated with copyright licensing.

Provide legal certainty

Legal certainty is fundamental for the effective extraction of biodiversity data from scholarly publications, their subsequent mobilization and reuse for science to address societies' challenges and for the conservation of biodiversity. The recommendations reduce the prospective transaction costs associated with copyright licensing. They do so by removing too-narrowly scoped copyright-specific obligations, which are already fulfilled by adhering to scientific best practices and community norms.

It is of importance for authors and publishers to be aware of and be able to differentiate between copyright protection of the publication itself as a creative work and the data within the publication. These are different elements that are independent of each other. Following our reasoning developed above, data present within publications are not copyrightable; however, the journal articles, books and publications of other types as a whole and in parts are private assets protected by copyright laws.

Existing publications often do not have clear copyright and license information associated with them. This can require intense background research into access authorization and reuse conditions for each of the publications from which data are to be reused. This is relevant for example, for research, digitization or data linking infrastructure projects. At the end of such inquiries into the legal status of a publication it is not uncommon that questions and uncertainties still remain.

Even if the legal conditions associated with publications are easy to find and clearly stated, the utilization of several different sources to assemble compound datasets can result in license stacking. Individual source publications and datasets might fall under a wide range of (national) copyright contexts and license statements, involving separate rights and legal obligations relating to publishers, authors and users. Specifically in investigations that are utilizing numerous resources from multiple, divergent

scientific backgrounds and including a range of data types with distinct cultures attached, this can create a patchwork of distinct and divergent conditions of use, which are hardly navigable for researchers assembling larger datasets.

To circumvent these difficulties and uncertainties, we suggest that authors and publishers as copyright holders explicitly and voluntarily choose to waive known or potentially applicable copyright to their works (as authors) or the publications in their portfolios (as publishers). This can be achieved by associating the Creative Commons copyright waiver (CC0) that acts as public domain dedication ("no rights reserved", https://creativecommons.org/public-domain/cc0/). Alternatively, copyright holders might choose a common-use-license to govern the reuse of their work. This can be achieved by applying one of the Creative Commons licenses, preferably the less restrictive CC-BY license ("attribution", https://creativecommons.org/licenses/by/4.0/legalcode.en). The Creative Commons CC-BY license enables open access and only requires attribution for reuse. At a minimum, we recommend that authors and publishers at least explicitly state the conditions for access and reuse of a work (e.g., Bouchout Declaration on Open Biodiversity Knowledge Management, Anonymous 2014). This can be achieved by voluntarily setting a work under the governance of a unilateral declaration (e.g., a common-use-license as suggested).

Both options, the CC0 waiver and the CC-BY copyright license, provide reasonable legal security. This is specifically of importance when it is unclear if the tangible work or product that an author created falls under copyright law of some jurisdictions. No matter the background and context, CC0 and CC-BY will ensure easy, clear and unambiguous access to and reusability of creative works and, hence, the data within them, by everyone including scientists from all backgrounds who employ TDM for data mobilization and linking.

Important cultural collection facilities and data aggregators use the CC0 public domain dedication, for example, the EU-funded digital library Europeana uses CC0 for its metadata since 2012 (https://creativecommons.org/2012/09/12/europeana-releases-20-million-records-into-the-public-domain-using-cc0/). Since 2017, the Metropolitan Museum of Art in New York City (NY, U.S.) sets its metadata as well as images that are in the public domain or to which the organization waives any copyright it might have into the public domain using CC0. Their goal is to make their resources "… widely and freely available for unrestricted use, and at no cost…" (see Open Access Policy at https://www.metmuseum.org/policies/image-resources).

The terms of use of Europeana (https://www.europeana.eu/en/rights/public-domain-usage-guidelines) include public domain usage guidelines (Appendix 1), which closely correspond to the data governance framework that we advocate here for the scientific and biodiversity data communities. This spirit forms the foundation for our argumentation and recommendations. Its narrative sustains our recommendation for rights holders of copyrightable works to set their publications into the public domain, the reuse of non-copyrightable data, and our proposal to develop sociocultural and ethical data governance labels.

We acknowledge that copyright protection fulfills an important role to protect the business foundation of creators and publishers. It might not be possible for all publishing houses to immediately waive copyright and/or provide open access to all of the creative works in their portfolio. A balance has to be found between public interests in open and free access to and reuse of a scientific work, and the need of especially smaller publishing houses, e.g., associated with professional societies, to protect the copyrightable components of their portfolio as a basis for generating income in support of business activities, e.g., to cover publishing costs, and thus retain their strategic autonomy. If a public domain dedication is not possible from the start, we recommend that a compromise should be found that protects the business interests of creators and publishers while keeping the needs of users, that is societies, in mind.

Mark data explicitly as public good

A key insight that our investigation highlights is that data cannot be copyrighted. Scientific data are public goods that legally are situated in the public domain. Once authorization for lawful access to a copyrighted scientific work has been obtained, the data within it can be openly accessed, detected, extracted, deposited elsewhere and then reused freely without any copyright-associated restrictions.

The significant value of such open research data to societies has been officially recognized both in the EU (Directive (EU) 2019/1024, European Parliament and Council 2019b) and the U.S. (Office of Science and Technology Policy 2022; see also Brainard & Kaiser 2022; National Science Foundation 2023). Both aim to promote its open availability and reuse. For example, Directive (EU) 2019/1024 (European Parliament and Council 2019b) requires "publicly funded research data" to be "open by default" (Art. 10 n. 1). The directive addresses required constraints to openness by stating that:

> "… concerns relating to intellectual property rights, personal data protection and confidentiality, security and legitimate commercial interests, shall be taken into account in accordance with the principle of 'as open as possible, as closed as necessary'." (European Parliament and Council 2019b Art. 10 n. 1).

However, no definitions or details are provided as to the applicability of intellectual property rights (which include copyright), that is "what" can be copyrighted (see above), how a status as open data might be communicated to users and which strategies are preferable to data providers and users for navigating landscapes that include both copyrighted works and open data. Our publication has the aim to close that gap and provide guidance.

Originally a work-around to declare that data is open, over the past two decades it has become a common procedure to attach open copyright licenses to data as if they were copyrightable. This habit, however, is confusing the overall situation for providers and users alike. Moreover, a better solution exists by now. For example, currently GBIF only accepts and thereby enforces the use of the CC0, CC-BY or CC-BY-NC (i.e., waiver, attribution and non-commercial, respectively) licenses (https://www.gbif.org/terms and https://ipt.gbif.org/manual/en/ipt/latest/license). Yet, these licenses in fact restrict the reuse of natively unrestricted data. For entities that are already in the public domain, it is more appropriate to use a public domain mark (PDM), for example the one provided by Creative Commons (https://creativecommons.org/public-domain/pdm/). A PDM provides the functionality to explicitly state that a certain entity, i.e., work or data, is known or considered to be in the public domain in all jurisdictions worldwide. While Creative Commons states that its PDM "is not legally operative in any respect – it is intended to function as a label" (https://wiki.creativecommons.org/wiki/PDM_FAQ), the PDM clearly conveys the intention of a data provider to make their resources freely available as public goods.

Reichman & Uhlir (2003: 422, footnote 538, also p. 319 and 331) point out the importance of proactively and expressively declaring and contextualizing data into the public domain. They argue that this will ensure that data will not be captured and made proprietary in any way in the future. Marking data as public goods by associating them with a PDM is a proactive step in preserving and strengthening the public domain and governing it by the digital public commons. Therefore, we recommend that authors and/or publishers highlight that the data within their published work are not copyrightable by marking the contents of publications as open and freely reusable resources that reside within the digital public domain.

Provide full attribution and transparency

We consider that a clear statement of provenance, including attribution of previous work to authors and their contributing collaborators, is a scientific norm, which is part of best practices and social codes of conduct fostering social capacity within the scientific community.

This standpoint seems in line with codes of conduct published by national science foundations, for example, the "Guidelines for Safeguarding Good Research Practice: Code of Conduct" by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft 2022). The DFG considers it part of a scientific quality assurance process for "Researchers [to] provide full and correct information about their own preliminary work and that of others" (p. 17, "Guideline 13: Providing public access to research results") and "… to ensure that citations are clear, and, as far as possible, to enable third parties to access this information" (p. 17, "Guideline 12: Documentation").

Good science requires transparency about the origin of ideas, concepts, hypotheses and data, and thus attribution. For example, only when a source is given, data retain a link to the methods and conditions under which they have been gathered – information and context that are crucial for interpreting results drawn from the data, and for developing follow-up hypotheses. Thus, attribution and provenance are essential for making it possible to evaluate new research and its outcomes. Being scientists, we attribute, cite and describe, not because we are legally bound to do so under the license terms, but because this approach results in effective and exciting data, hypothesis tests and conclusions that can be reproduced. As scientists, we are also aware that we are standing on the shoulders of giants, all the bright, enthusiastic and inquisitive individuals who came before us and have explored their worlds, both outside and inside.

Be explicit about data governance

Science's practices and codes are based on an understanding that data are not owned, but represent a common achievement, to be made openly and freely accessible and available, and to be shared and reused for fostering scientific inquiry and progress as contributions to the public good (Kalkman *et al.* 2019; Salwen 2021). This is recognized by recommendations and guidelines, for example, the *Bouchout Declaration on Open Biodiversity Knowledge Management* (Anonymous 2014), the *Legal Interoperability of Research Data: Principles and Implementation Guidelines* (RDA-CODATA Legal Interoperability Interest Group 2016), the *GEO Statement on Open Knowledge* (Group on Earth Observations 2021), the *Recommendation on Open Science* (UNESCO 2021), the *Recommendation of the Council on Enhancing Access to and Sharing of Data* (OECD 2021), but also legislative works such as Directive (EU) 2019/1024 of the European Parliament and Council (2019b).

Nevertheless, in addition to being best scientific practice and providing transparency as discussed previously, attribution of works and data to scientists that contribute to the research process is essential for research assessments and, hence, of direct interest and value to data and service providers. These encompass, for example, scientists and their research institutions, as well as data repositories, aggregators and associated organizations (e.g., professional societies and standards organizations). Provenance and attribution provide visibility to efforts, investments, engagement and collaborations for research and endeavors of finding solutions to societal challenges. In evaluation processes, e.g., by funders, citations are quantified and citation patterns are used to qualify range, connectivity and impact.

Thus, it is crucial for providers of data, authors and publishers of scholarly publications and standard developers to express their needs and wishes (e.g., for attribution and visibility), to have their own and outside preferences and pressures met. We recommend to complement explicit statements of the open and free reusability of works (CC0) and data (PDM) with a clear and standardized notice in the governance section of a publication or data repository submission that states the author(s)' culture, needs, expectations and preferences. Such governance notices, for example, can inform users about a specific need for attribution and/or if the author(s) are interested in open dialogues and collaboration.

Local Contexts labels (https://localcontexts.org/) developed by Indigenous Peoples and Local Communities (IPLCs) in collaboration with scientists communicate authors' and providers' backgrounds, needs and preferences. In our view, open data, open science and the digital public commons in the science sector require an equivalent and the Local Contexts labels provide a suitable starting point for further development. "[T]he Labels allow communities to express local and specific conditions for sharing and engaging in future research and relationships" (https://localcontexts.org/labels/traditional-knowledge-labels/). The labels are to be associated with traditional knowledge, as well as biocultural collections and data, respectively, and have thematic groups of labels focusing on provenance, protocols and permissions.

Label-based approaches are successfully applied in several contexts. The Local Contexts labels of IPLCs are similar in function to the Creative Commons public domain mark (PDM), which also has the status of a label, informing users that a work or data are considered to be in the public domain worldwide. A label-based approach is also taken by cultural heritage institutions to associate and communicate standardized rights statements with their online cultural heritage records (https://rightsstatements.org/en/). It would be equally appropriate and effective for the scientific community to develop sociocultural and ethical labels for expressing and disseminating applicable contexts in which data and providers are situated, communicating providers' needs and expectations as well as for reaching out and communicating providers' interests.

In the biodiversity sciences, TDWG (https://www.tdwg.org/) as a not-for-profit organization dedicated to developing and maintaining information standards would be well-positioned to develop such sociocultural and ethical labels. Its dense global network of community members and organizations has the expertise and protocols in place to develop such labels. It can find agreement for their community-wide ratification, and could set into practice machine-actionable labels that express the contexts in which data and works exist as outcomes of the engagement of the biodiversity research and collections communities.

Mark-up your work in support of machine-actionable data mobilization

We have clarified the (non-)copyrightability of data and described a process to extract them from copyrightable publications, developed a robust approach to dedicate data and publications as open resources into the digital public commons and to communicate sociocultural and ethical governance conditions. Our propositions pave the way for extraction and reuse of data from scholarly publications. What remains are social, more than technical, obstacles that need to be overcome.

We recommend that publishers mark-up all parts and contents of manuscripts during the publishing process. Such mark-up will transform manuscripts into machine-actionable resources that are enabling, improving and enhancing efficient and large-scale TDM workflows for the generation and mobilization of FAIR data and knowledge.

Again, it has to be stressed that TDM does not affect the copyright status of creative works that contain scientific data. Journals, journal articles, books, book chapters and other types of publications as a whole are and remain intellectual property assets protected by copyright laws. Therefore, the business foundation of publishers in the form of their portfolios would not be affected by large-scale, more frequent and widespread TDM. Moreover, with data becoming associated with comprehensive provenance information as a global community standard and subsequently automated and machine-actionable attribution modules being established, much increased reuse of mobilized data will in turn lead to enhanced citation of scholarly publications and hence increase visibility of publishers' portfolios and their value.

Conclusions and outlook

This work provides the development of arguments and reasoning underlying the recently published joint recommendation by BHL, CETAF and SPNHC (Bénichou *et al.* 2023) on the (non-)copyrightability of data within scholarly publications.

To strengthen and expand the digital public commons that enables open science, we recommend voluntarily waiving copyright to publications (e.g., CC0) or using an open copyright license with a condition for attribution (e.g., CC-BY) for governing access to publications and the data within them. Having determined that the data themselves are not copyrightable, we recommend explicitly declaring that they are a public good through the use of a public domain mark (e.g., the PDM by Creative Commons). In this way, open and free access to and reuse of data is safeguarded into the future. Machine-actionable markup technologies structuring and formatting the digital versions of publications will provide the technical basis for efficient and effective large-scale TDM and thus data mobilization.

The outcome of our investigation into the legal context of copyright laws clarifies the legal aspects and describes sociocultural and ethical as well as data governance considerations associated with accessing and reusing data from scholarly works. Community norms and scientific best practices are of fundamental importance and lie at the core of solutions with real-world applicability.

Adaptation and continuous development of existing best practice guidelines in use by the biodiversity sciences and informatics community will be key for evolving the legal landscape and the global policy contexts for digital data, information and knowledge. Clarity and legal certainty will accelerate data reuse, which will be essential for addressing the scientific priorities of society, innovation and progress.

Acknowledgements

This joint investigation was initiated by Laurence Bénichou, co-chair of the CETAF ePublishing working group. It brought together and combined several independent work streams, including CETAF ePublishing working group's work on best practices in publishing, the participation of CETAF and SPNHC members in the alliance for biodiversity knowledge's consultation for the Digital Extended Specimen (DES) concept (Hardisty *et al.* 2022) in 2021, science-policy advocacy and partnership building with the UN Convention on Biological Diversity by SPNHC's Biodiversity Crisis Response Committee, and BHL long-standing initiatives in making freely and openly available biodiversity works and data. A concise companion paper focused on the four recommendations that was published recently was approved by the executive boards of CETAF, SPNHC and the BHL and thus received the support of the three organisations.

The text was much improved by the comments of several colleagues. We would like to very much thank Gergely Babocsay, David Iggulden, Martin Kalfatovic, Anahita Kazem, Jiří Kvaček, Michelle Price and Scott Rufolo.

The authors would like to give special recognition to the late Constance Rinaldo, Washington D.C., U.S. (https://orcid.org/0000-0002-8339-728X), a member of the Biodiversity Heritage Library. Connie was a founding member of the group, together with Laurence Bénichou, Donat Agosti and Jutta Buschbom. Her engagement moved the initial process forward, and she was contributing to the preparation of the joint-workshop at the TDWG meeting in 2022 before she passed away on October 27, 2022. It was our aim to continue the development of the workshop and its resulting publications in Connie's spirit.

Last but not least, we express our gratitude to the two reviewers for their invaluable comments and suggestions that helped to evolve the document.

Authors' contributions

LB, DA and WE conceived the original concept recognizing existing needs in the collections-based community and formulated the recommendations. Joining them, JB, MK, PM, AK and EH extended the investigation into clarifying underlying rationales and evolving a general data governance perspective. WE and MK as lawyers with a focus on copyright and open science contributed legal expertise. With professional backgrounds in the publishing and library sectors, LB and EH provided expertise and experiences as providers and users from these sectors. AK and DA are providers of pipelines and infrastructures for open data, PM and JB are engaged in science-policy advocacy with a focus on governance of open data. All authors contributed substantially to the development of the scope, argumentation and proposed solutions. LB and JB took the lead in writing the manuscript. DA and WE restructured the manuscript following the first round of reviews. All authors provided extensive input to the manuscript and critical feedback to its contents.

Views expressed and legal disclaimers

The views expressed in this publication are those of the authors and do not reflect the official positions or legal opinions of authors' affiliations or employers. The authors have contributed to this publication in their personal capacity only.

The statements in this publication represent scientific points of view. They do not constitute legal advice.

References

Adriaans P. 2023. Information. *In*: Zalta E.N. & Nodelman U. (eds) *The Stanford Encyclopedia of Philosophy (Winter 2023 Edition)*. Metaphysics Research Lab, Stanford University, Stanford, CA, USA. Available from https://plato.stanford.edu/archives/win2023/entries/information/.

Agosti D., Bénichou L., Addink W., Arvanitidis C., Catapano T., Cochrane G., Dillen M., Döring M., Georgiev T., Gérard I., Groom Q., Kishor P., Kroh A., Kvaček J., Mergen P., Mietchen D., Pauperio J., Sautter G. & Penev L. 2022. Recommendations for use of annotations and persistent identifiers in taxonomy and biodiversity publishing. *Research Ideas and Outcomes* 8: e97374. https://doi.org/10.3897/rio.8.e97374

Anonymous 2014. The Bouchout Declaration for Open Biodiversity Knowledge Management. Available from https://www.bouchoutdeclaration.org/ [accessed 24 Jun. 2025].

Ball A. 2014. How to license research data. Digital Curation Centre: DCC How-to Guides. Available from https://www.dcc.ac.uk/guidance/how-guides/license-research-data [accessed 24 Jun. 2025].

Bénichou L., Gérard I., Laureys É. & Price M. 2018. Consortium of European Taxonomic Facilities (CETAF) best practices in electronic publishing in taxonomy. *European Journal of Taxonomy* 475: 1–37. https://doi.org/10.5852/ejt.2018.475

Bénichou L., Guidoti M., Gérard I., Agosti D., Robillard T. & Cianferoni F. 2021. European Journal of Taxonomy: a deeper look into a decade of data. *European Journal of Taxonomy* 782: 173–196. https://doi.org/10.5852/ejt.2021.782.1597

Bénichou L., Buschbom J., Campbell M., Hermann E., Kvacek J., Mergen P., Mitchell L., Rinaldo C. & Agosti D. 2022. Joint statement on best practices for the citation of authorities of scientific names in taxonomy by CETAF, SPNHC and BHL. *Research Ideas and Outcomes* 8: e94338. https://doi.org/10.3897/rio.8.e94338

Bénichou L., Agosti D., Egloff W., Hermann E., Kageyama M., Mergen P., Rinaldo C. & Buschbom J. 2023. Joint statement by CETAF, SPNHC and BHL on DATA within scientific publications: clarification of [non]copyrightability. *Research Ideas and Outcomes* 9: e115466. https://doi.org/10.3897/rio.9.e115466

Brainard J. & Kaiser J. 2022. U.S. to require free access to papers on all research it funds. *Science* 377 (6610): 1026–1027. https://doi.org/10.1126/science.ade6577

Bundesversammlung der Schweizerischen Eidgenossenschaft. 1992. SR 231.1: Bundesgesetz vom 9. Oktober 1992 über das Urheberrecht und verwandte Schutzrechte (Urheberrechtsgesetz, URG) (Federal Act on Copyright and Related Rights (Copyright Act, CopA) of 9 October 1992) (Status as of 1 July 2023). Available from https://www.fedlex.admin.ch/en/cc/internal-law/23#231.

Carroll S.R., Garba I., Figueroa-Rodríguez O.L., Holbrook J., Lovett R., Materechera S., Parsons M., Raseroka K., Rodriguez-Lonebear D., Rowe R., Sara R., Walker J.D., Anderson J. & Hudson M. 2020. The CARE Principles for indigenous data governance. *Data Science Journal* 19: 43. https://doi.org/10.5334/dsj-2020-043

Carroll S.R., Herczog E., Hudson M., Russell K. & Stall S. 2021. Operationalizing the CARE and FAIR Principles for indigenous data futures. *Scientific Data* 8 (1): 108. https://doi.org/10.1038/s41597-021-00892-0

CBD. 2022. 15/4. Kunming-Montreal Global Biodiversity Framework (CBD/COP/DEC/15/4). Available from https://www.cbd.int/conferences/2021-2022/cop-15/documents.

Coalition for Advancing Research Assessment. 2024. About. Available from https://coara.eu/ [accessed 24 Jun. 2025].

Cosens B., Ruhl J.B., Soininen N., Gunderson L., Belinskij A., Blenckner T., Camacho A.E., Chaffin B.C., Craig R.K., Doremus H., Glicksman R., Heiskanen A.-S., Larson R. & Simila J. 2021. Governing complexity: Integrating science, governance, and law to manage accelerating change in the globalized commons. *Proceedings of the National Academy of Sciences of the United States of America* 118 (36): e2102798118. https://doi.org/10.1073/pnas.2102798118

Council of the European Union. 2022. Research assessment and implementation of Open Science, 10126/22. Available from https://www.consilium.europa.eu/media/56958/st10126-en22.pdf.

Council of the European Union. 2023. High-quality, transparent, open, trustworthy and equitable scholarly publishing, 9616/23. Available from https://data.consilium.europa.eu/doc/document/ST-9616-2023-INIT/en/pdf.

Deutsche Forschungsgemeinschaft. 2022. Guidelines for Safeguarding Good Research Practice: Code of Conduct. Available from https://doi.org/10.5281/zenodo.6472827.

Deutscher Bundestag. 2021. Urheberrechtsgesetz vom 9. September 1965 (BGBl. I S. 1273), das zuletzt durch Artikel 25 des Gesetzes vom 23. Juni 2021 (BGBl. I S. 1858) geändert worden ist (UrhG) (German federal copyright law from September 9, 1965, changed most recently on June 23, 2021). Available from https://www.gesetze-im-internet.de/urhg/BJNR012730965.html#BJNR012730965BJNG004800123.

Dulong de Rosnay M. & Stalder F. 2020. Digital commons. *Internet Policy Review* 9 (4). https://doi.org/10.14763/2020.4.1530

Egloff W., Agosti D., Patterson D., Hoffmann A., Mietchen D., Kishor P. & Penev L. 2016. Data policy recommendations for biodiversity data. EU BON Project Report. *Research Ideas and Outcomes* 2: e8458. https://doi.org/10.3897/rio.2.e8458

Egloff W., Agosti D., Kishor P., Patterson D. & Miller J. 2017. Copyright and the use of images as biodiversity data. *Research Ideas and Outcomes* 3: e12502. https://doi.org/10.3897/rio.3.e12502

European Commission. 2020. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions – A European strategy for data. COM(2020)66. Available from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066.

European Commission. 2021. Towards a reform of the research assessment system: scoping report. Publication Office. Available from https://data.europa.eu/doi/10.2777/707440.

European Commission Directorate-General for Research and Innovation & Angelopoulos C. 2022. Study on EU copyright and related rights and access to and reuse of scientific publications, including open access – Exceptions and limitations, rights retention strategies and the secondary publication right. Available from https://data.europa.eu/doi/10.2777/891665.

European Parliament and Council. 1996. Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases. Available from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31996L0009.

European Parliament and Council. 2019a. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Text with EEA relevance.). Available from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32019L0790.

European Parliament and Council. 2019b. Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast). Available from https://eur-lex.europa.eu/eli/dir/2019/1024/oj.

European Parliament and Council. 2022. Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act). Available from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R0868.

European Parliament and Council. 2023. Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act). Available from https://eur-lex.europa.eu/eli/reg/2023/2854.

Gervais D.J. 2017. *(Re)structuring Copyright*. Edward Elgar Publishing, Cheltenham, UK / Northampton, MA, USA. Available from https://www.e-elgar.com/shop/gbp/re-structuring-copyright-9781789902143.html.

GO FAIR. 2024. FAIRification Process. Available from https://www.go-fair.org/fair-principles/fairification-process/ [accessed 24 Jun. 2025].

Group on Earth Observations. 2021. GEO Statement on Open Knowledge. Available from https://earthobservations.org/storage/documents/Open-Knowledge/GEO-17-4.1_GEO%20Statement%20on%20Open%20Knowledge.pdf [accessed 24 Jun. 2025].

Guidi S. 2023. Innovation commons for the data economy. *Digital Society* 2 (2): 31. https://doi.org/10.1007/s44206-023-00059-x

Hahnel M., Smith G., Schoenenberger H., Scaplehorn N. & Day L. 2023. The State of Open Data 2023. *Digital Science Report*. Available from https://doi.org/10.6084/m9.figshare.24428194.v1.

Hardisty A., Ellwood E.R., Nelson G., Zimkus B., Buschbom J., Addink W., Rabeler R., Bates J., Bentley A., Fortes J.A.B., Hansen S., Macklin J.A., Mast A., Miller J.T., Monfils A.K., Paul D.L., Wallis E. & Webster M. 2022. Digital Extended Specimens: Enabling an extensible network of biodiversity data records as integrated digital objects on the Internet. *BioScience* 72 (10): 978–987. https://doi.org/10.1093/biosci/biac060

Hudson M., Carroll S.R., Anderson J., Blackwater D., Cordova-Marks F.M., Cummins J., David-Chavez D., Fernandez A., Garba I., Hiraldo D., Jager M.B., Jennings L.L., Martinez A., Sterling R., Walker J.D. & Rowe R.K. 2023. Indigenous peoples' rights in data: a contribution toward indigenous research sovereignty. *Frontiers in Research Metrics and Analytics* 8: 1173805. https://doi.org/10.3389/frma.2023.1173805

Huemer M.-A. 2021. Revision of Directive 96/9/EC on the legal protection of databases. Briefing – Implementation Appraisal. Available from https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)694232 [accessed 24 Jun. 2025].

Kalkman S., Mostert M., Gerlinger C., van Delden J.J.M. & van Thiel G.J.M.W. 2019. Responsible data sharing in international health research: a systematic review of principles and norms. *BMC Medical Ethics* 20 (1): 21. https://doi.org/10.1186/s12910-019-0359-9

Lin D., Crabtree J., Dillo I., Downs R.R., Edmunds R., Giaretta D., De Giusti M., L'Hours H., Hugo W., Jenkyns R., Khodiyar V., Martone M.E., Mokrane M., Navale V., Petters J., Sierman B., Sokolova D.V., Stockhause M. & Westbrook J. 2020. The TRUST Principles for digital repositories. *Scientific Data* 7 (1): 144. https://doi.org/10.1038/s41597-020-0486-7

Ministère de la Culture. 2021. Ordonnance n° 2021-1518 du 24 novembre 2021 complétant la transposition de la directive 2019/790 du Parlement européen et du Conseil du 17 avril 2019 sur le droit d'auteur et les droits voisins dans le marché unique numérique et modifiant les directives 96/9/CE et 2001/29/CE. Journal Officiel de la République Française (JORF) n° 0274 du 25 novembre 2021, Texte n° 15 (NOR : MICB2121839R). Available from https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000044362034.

Musgrave R.A. & Musgrave P.B. 1989. *Public Finance in Theory and Practice*. 5th (International) Edition. McGraw-Hill, New York, NY, USA.

National Science Foundation. 2023. NSF Public Access Plan 2.0: Ensuring Open, Immediate and Equitable Access to National Science Foundation Funded Research. National Science Foundation, Washington, D.C., USA. Available from https://www.nsf.gov/pubs/2023/nsf23104/nsf23104.pdf [accessed 24 Jun. 2025].

Nikander P., Eloranta V., Karhu K. & Hiekkanen K. 2020. Digitalisation, anti-rival compensation and governance: Need for experiments. Abstract from Nordic Workshop on Digital Foundations of Business, Operations, and Strategy. Espoo, Finland. Available from https://acris.aalto.fi/ws/portalfiles/portal/41477511/Nikander_et_al_2nd_DBOS.pdf [accessed 24 Jun. 2025].

OECD. 2021. Recommendation of the Council on Enhancing Access to and Sharing of Data, OECD/LEGAL/0463. Available from https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0463#mainText [accessed 24 Jun. 2025].

Office of Science and Technology Policy. 2022. Memorandum for the heads of executive departments and agencies: Ensuring free, immediate, and equitable access to federally funded research. Available from https://bidenwhitehouse.archives.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf [accessed 24 Jun. 2025].

Oldham P.D., Chiarolla C. & Thambisetty S. 2023. Digital Sequence Information in the UN High Seas Treaty: Insights from the Global Biodiversity Framework-related Decisions. *LSE Law School – Policy Briefing Series* 53. http://doi.org/10.2139/ssrn.4343130

Patterson D.J., Egloff W., Agosti D., Eades D., Franz N., Hagedorn G., Rees J.A. & Remsen D.P. 2014. Scientific names of organisms: attribution, rights, and licensing. *BMC Research Notes* 7: 79. https://doi.org/10.1186/1756-0500-7-79

Paul E.S. & Stokes D. 2023. Creativity. *In*: Zalta E.N. & Nodelman U. (eds) *The Stanford Encyclopedia of Philosophy (Spring 2023 Edition)*. Metaphysics Research Lab, Stanford University, Stanford, CA, USA. Available from https://plato.stanford.edu/archives/spr2023/entries/creativity/.

Penev L., Koureas D., Groom Q., Lanfear J., Agosti D., Casino A., Miller J., Arvanitidis C., Cochrane G., Hobern D., Banki O., Addink W., Kõljalg U., Copas K., Mergen P., Güntsch A., Bénichou L., Benito Gonzalez Lopez J., Ruch P., Martin C., Barov B., Demirova I. & Hristova K. 2022. Biodiversity Community Integrated Knowledge Library (BiCIKL). *Research Ideas and Outcomes* 8: e81136. https://doi.org/10.3897/rio.8.e81136

Platt J.R. 1964. Strong inference. *Science* 146 (3642): 347–353. https://doi.org/10.1126/science.146.3642.347

Plazi. 2023. 15 years of discovering known biodiversity. Available from http://plazi.org/posts/2023/12/15-years/ [accessed 24 Jun. 2025].

Purtova N. & van Maanen G. 2024. Data as an economic good, data as a commons, and data governance. *Law, Innovation and Technology* 16 (1): 1–42. https://doi.org/10.1080/17579961.2023.2265270

RDA-CODATA Legal Interoperability Interest Group. 2016. Legal Interoperability of Research Data: Principles and Implementation Guidelines. Available from https://doi.org/10.5281/zenodo.162241 [accessed 24 Jun. 2025].

Reichman J.H. & Okediji R.L. 2012. When copyright law and science collide: Empowering digitally integrated research methods on a global scale. *Minnesota Law Review* 96 (4): 1362–1480. Available from https://www.minnesotalawreview.org/wp-content/uploads/2012/08/ReichmanOkediji_MLR1362.pdf.

Reichman J.H. & Uhlir P.F. 2003. A contractually reconstructed research commons for scientific data in a highly protectionist intellectual property environment. *Law and Contemporary Problems* 66 (1): 315–462. Available from https://scholarship.law.duke.edu/lcp/vol66/iss1/12/.

Reichman J.H. & Uhlir P.F. 2004. A contractually reconstructed research commons for scientific data: International considerations. *In*: *Open Access and the Public Domain in Digital Data and Information for Science: Proceedings of an International Symposium*: 98–102. The National Academies Press, Washington, D.C., USA. Available from https://nap.nationalacademies.org/read/11030/chapter/25 [accessed 24 Jun. 2025].

Salwen H. 2021. Research ethical norms, guidance and the internet. *Science and Engineering Ethics* 27 (6): 67. https://doi.org/10.1007/s11948-021-00342-5

Sequoiah-Grayson S. & Floridi L. 2022. Semantic conceptions of information. *In*: Zalta E.N. (ed.) *The Stanford Encyclopedia of Philosophy (Spring 2022 Edition)*. Metaphysics Research Lab, Stanford University, Stanford, CA, USA. Available from https://plato.stanford.edu/archives/spr2022/entries/information-semantic/.

Smith A.M., Katz D.S. & Niemeyer K.E. 2016. Software citation principles. *PeerJ Computer Science* 2: e86. https://doi.org/10.7717/peerj-cs.86

U.S. Supreme Court. 1991. Feist Publications, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340. Available from https://tile.loc.gov/storage-services/service/ll/usrep/usrep499/usrep499340/usrep499340.pdf.

UNESCO. 2021. UNESCO Recommendation on Open Science. Available from https://doi.org/10.54677/MNMH8546.

UNGA. 2020. Road map for digital cooperation: implementation of the recommendations of the High-level Panel on Digital Cooperation. Report of the Secretary-General. Available from https://www.un.org/en/content/digital-cooperation-roadmap/ [accessed 24 Jun. 2025].

Office of the Law Revision Counsel of the House of Representatives. 2023a. United States Code, 2018 Edition, Supplement 5, Title 17 – Copyrights. Chapter 1 – Subject matter and scope of copyright (Sections 101-122), § 106. Exclusive rights in copyrighted works. Available from https://www.govinfo.gov/app/collection/uscode/2023/title17/chapter1.

Office of the Law Revision Counsel of the House of Representatives. 2023b. United States Code, 2018 Edition, Supplement 5, Title 17 – Copyrights. Chapter 1 – Subject matter and scope of copyright (Sections 101-122), § 107. Limitations on exclusive rights: Fair use. Available from https://www.govinfo.gov/app/collection/uscode/2023/title17/chapter1.

Watanabe M.E. 2018. Digitizing specimens - Legal issues abound. *Bioscience* 68 (9): 728–728. https://doi.org/10.1093/biosci/biy086

Wilkinson M.D., Dumontier M., Aalbersberg I.J.J., Appleton G., Axton M., Baak A., Blomberg N., Boiten J.-W., da Silva Santos L.B., 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., 't Hoen P.A.C., Hooft R., Kuhn T., Kok R., Kok J., Lusher S.J., Martone M.E., Mons A., Packer A.L., Persson B., Rocca-Serra P., Roos M., van Schaik R., Sansone S.-A., Schultes E., Sengstag T., Slater T., Strawn G., Swertz M.A., Thompson M., van der Lei J., van Mulligen E., Velterop J., Waagmeester A., Wittenburg P., Wolstencroft K., Zhao J. & Mons B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3: 160018. https://doi.org/10.1038/sdata.2016.18

WIPO. 1979. Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979). TRT/BERNE/001. Available from https://www.wipo.int/treaties/en/ip/berne/index.html.

Printed versions of all papers are deposited in the libraries of four of the institutes that are members of the *EJT* consortium: Muséum national d'Histoire naturelle, Paris, France; Meise Botanic Garden, Belgium; Royal Museum for Central Africa, Tervuren, Belgium; Royal Belgian Institute of Natural Sciences, Brussels, Belgium. The other members of the consortium are: Natural History Museum of Denmark, Copenhagen, Denmark; Naturalis Biodiversity Center, Leiden, the Netherlands; Museo Nacional de Ciencias Naturales-CSIC, Madrid, Spain; Leibniz Institute for the Analysis of Biodiversity Change, Bonn – Hamburg, Germany; National Museum of the Czech Republic, Prague, Czech Republic; The Steinhardt Museum of Natural History, Tel Aviv, Israel.

Appendix

Appendix 1. Europeana's guidelines for the use of data that is in the public domain (https://www.europeana.eu/en/rights/public-domain-usage-guidelines) state the following best practice rules:

- Give credit where credit is due.
- Protect the reputation of creators and providers.
- Show respect for the original work.
- Show respect for the creator.
- Share knowledge.
- Be culturally aware.
- Support efforts to enrich the public domain.
- Preserve public domain marks and notices.
- This usage guide is based on goodwill.