Joint statement by CETAF, SPNHC and BHL on DATA within scientific publications: clarification of [non]copyrightability

Laurence Benichou[‡,§], Donat Agosti[|,‡], Willi Egloff[|,‡], Elisa Hermann[¶,#], Mariko Kageyama[¤,«], Patricia Mergen[»,^,‡], Constance Rinaldo[#], Jutta Buschbom[˅,¦,«]

‡ CETAF E-Publishing Working group, Brussels, Belgium
§ Museum national d'Histoire naturelle, Paris, France
| Plazi, Bern, Switzerland
¶ Museum für Naturkunde - Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany
# Biodiversity Heritage Library, Washington D.C., United States of America
¤ Independent Consultant, Seattle, Washington, United States of America
« Society for the Preservation of Natural History Collection, Chicago, United States of America
» Meise Botanic Garden, Meise, Belgium
^ Royal Museum for Central Africa, Tervuren, Belgium
˅ Statistical Genetics, Ahrensburg, Germany
¦ Natural History Museum, London, United Kingdom

Corresponding author: Laurence Benichou (laurence.benichou@mnhn.fr)

Abstract

The EU and other states have made legislative efforts to clarify data mining in copyrightable works, but the situation remains obscure and confusing, especially in a globalised field where international legislation can contribute to opacity. The present paper aims at asserting a common position of three communities representing biodiversity sciences and data specialists on this issue and to propose common and best practice guidelines so that they become universally accepted rules.

As scientific data users, we take the standpoint that scientific data are not copyrightable and, furthermore, they can be accessed, shared and reused freely. Thus, once legal access has been gained to copyrighted publications, the data within those scholarly publications can be considered to be open data that is freely extractable. This set of recommendations has been reached specifically for scientific use and societal benefits.

Keywords

copyright, data, text and data mining, taxonomic work, biodiversity, scientific publications

Introduction

This paper is the outcome of a workshop organised in October 2022 during the annual meeting of TDWG, the Biodiversity Information Standards organisation, held in Sofia, Bulgaria. The workshop was jointly organised by members of the Biodiversity Heritage Library (BHL), the e-Publishing working group of the Consortium of European Taxonomic Facilities (CETAF) and the Society for the Preservation of Natural History Collections (SPNHC) and supported by the Biodiversity Community Integrated Knowledge Library (BiCIKL; Penev et al. (2022)) project. The focus of the workshop was on the legal and contractual rules governing data within copyrighted works. The goals of its recommendations are to empower the biodiversity sciences and data community, including publishers, authors and users, to use appropriate legal and contractual licences and language that will allow data to be reused. A more in-depth discussion is provided by the authors (Buschbom in press). Building on this, the aim is to develop a common vision and a way forward that will allow and accelerate the extraction and reuse of data contained within publications, both legacy and prospective.

Clarifying the legal, ethical and socio-cultural contexts of FAIR (Findable, Accessible, Interoperable and Reusable) data (Wilkinson et al. 2016), we recommend a set of best practices that provide legal clarity, as well as attribution, transparency and accountability for the extraction and reuse of often high quality and information-rich biodiversity data from copyrighted works, specifically scholarly publications. Such data can be integrated into the body of the publication itself, for example, in the form of free text, tables, images or identification keys or attached to it as supplementary datasets.

The proposed set of recommendations builds on existing frameworks, as for example, the Bouchout Declaration on Open Biodiversity Knowledge Management (Anonymous 2014), 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). This set of recommendations considers existing discussions of copyright-associated questions in scientific contexts (e.g. Watanabe 2018; European Commission, Directorate-General for Research and Innovation and Angelopoulos 2022) The proposed recommendations reinforce existing best practice guidelines (Ball 2014; Patterson et al. 2014; Egloff et al. 2016, Egloff et al. 2017; Bénichou et al. 2018, Bénichou et al. 2021, Benichou et al. 2022) in use by the biodiversity sciences and informatics community and adapts them to the evolving legal landscape and changing global policy contexts of the ongoing digital transformation.

Description of the problem

Currently, most small publishers, specifically institutional 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 procedures, even unmanageable, for extracting data from publications widely persist, negatively affecting the productivity of biodiversity scientists and data managers who are interested in, and dependent on, the re-use 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.

While scientific publications, by default, are works protected by copyright, **scientific data are not copyrightable.** Their form is dictated by applicable standards, technical capacity and scientific good practices, which means that data in themselves are neither the result of creative choices nor expressive elements of a work made by the author(s). Furthermore, the copyright protection of a publication refers to the work, not to the data contained in it (499 U.S. 340 1991, Feist vs. Rural, U.S. Supreme Court 1991; Gervais 2019).

Liberating data from existing publications therefore means – from a copyright point of view – extracting unprotected data from protected works, often referred to as text and data mining. We understand text and data mining as "any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes, but is not limited to patterns, trends and correlations", as defined in Art. 2 n. 2 EU Directive 790/2019 (European Parliament and Council 2019). As this automated procedure includes the reuse of the protected work (as do some manual approaches as well), access to and reuse of the work needs an authorisation. This authorisation can be given by contractual licence or by legal licence. Legal licences 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).

The EU Directive 790/2019 has introduced two compulsory legal licences referring to text and data mining: Art. 3 obliges every Member State to introduce into its national copyright law a compulsory legal licence for text and data mining for the purposes of scientific research conducted by recognised research organisations and cultural heritage institutions. Art. 4 obliges them to introduce a subsidiary legal licence for any form of text and data mining for any other purpose.

As a result, copyright legislation actually presents a legal divide: in the EU, extracting data from publications for the purposes of scientific research is allowed by law. This authorisation prevails over any contractual agreement and also over eventual licences (as for example CC-licences). In the US, the same procedure may require a contractual

licence, unless the conditions for “fair use” are satisfied. In the rest of the world, the legislation differs from country to country.

In Switzerland, extracting data from publications is allowed by legal licence since a revision of the Swiss copyright law in 1992 (SR 231.1 1992). This is why Plazi has based its extraction workflow in Switzerland. Systematic extraction of taxonomic data from scientific publications started in 2009. Since 2013, the extracted data have been deposited in the Biodiversity Literature Repository in Zenodo, a general-purpose open repository developed under the European OpenAIRE programme and operated by CERN (Conseil européen pour la Recherche nucléaire). There has never been any dispute referring to an alleged copyright infringement.

Beyond copyright, it is good scientific practice to attribute extracted data to the source of extraction (Wilkinson et al. 2016; EOSC 2023). Once legally extracted, data can be reused freely. Some restrictions may apply from other protection schemes such as those concerning the protection of national security, the right of privacy and the protection of endangered species. However, we would point out that attribution and credit should not be confused with copyright. From a copyright point of view, extracted data can be reused worldwide without further authorisation.

As with existing legacy publications and data contained within them, it is important for authors and publishers to be aware of the legal situation and the differentiation between the copyright concerning the publication as a whole and copyright of the data within it, as these are matters that are independent of each other.

Thus, journal articles and books as a whole are and remain assets protected by copyright laws and regulations. Therefore, the business foundation of publishers and the business intelligence represented by their portfolios is not affected by the recommendations presented below. These consider solely the scientific data present in the publications.

Experiences with existing publications and data contained in them demonstrate that they often do not have clear copyright and licence information enabling and supporting reuse associated with them. This can require intense background research for each publication about which data are to be used within a research, digitisation or data interlinking infrastructure project. At the end of such inquiries into the legal status, it is not uncommon that questions and uncertainties still remain.

Even if the legal conditions associated with publications and data within them are easily accessible and clearly stated, specifically in investigations utilising many resources from multiple, divergent scientific backgrounds and including various data types, the individual source publications and their data might fall under a wide range of (national) copyright contexts and licence statements implicating the rights and obligations of publishers, authors and users. This creates a patchwork of distinct and divergent conditions, which are difficult to navigate for researchers assembling large datasets.

Looking forward, a solution to the current often ambiguous and patchy situation in the publishing landscape is to explicitly designate scientific data within publications as open

and freely reusable, which will result in harmonisation and increased availability of machine-actionable data.

Recommendations

The proposed set of recommendations focuses on the copyright law aspects and scientific best practice norms for accessing and reusing data from scholarly works. The recommendations clarify and adapt existing best practice guidelines in use by the biodiversity sciences and informatics community to the evolving legal landscape and changing global policy contexts for digital information, as well as data needs for answering today's challenges. As societies and associations, we recommend that:

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

2. authors and publishers explicitly state that they consider scientific data as not copyrightable. Best practice is to set the contents of their publications, be it data, drawings, media objects etc. (see the Blue List below) into the public domain by attaching a public domain mark that provides certainty about their reusability;

3. publishers use a publishing technique 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 and — for works having more than one author — the respective contributions of all co-authors.

It is best practice in scientific communities to work on the basis of scientific norms that exist independently of the legal realm with its laws, regulations, licences and agreements. These scientific norms exist in the form of well-established best practice approaches to scientific processes and an overarching community code of conduct. Our practices and codes state 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; Salwén 2021). Data sharing and its associated comprehensive attribution form an important component of the unwritten though widely agreed norms, practices and codes that are in place for fostering transparency, reproducibility and accountability.

As a wider scientific community, it is important to reiterate that the data contained in a scientific publication are freely extractable and reusable. This holds true, in particular, for those parts of the text that form the basis of a taxonomic treatment, as formerly described in the Blue List established by Patterson et al. (2014) and updated here:

1. A hierarchical organisation (classification), in which, as examples, species are nested in genera, genera in families, families in orders and so on;

2. Alphabetical, chronological, phylogenetic, palaeontological, geographical, ecological, host-based or feature-based (e.g. life-form) ordering of taxa;

3. Scientific names of genera or other uninomial taxa, species, epithets of species names, binomial combinations as species names or names of infraspecific taxa; with or without the author of the name and the date when it was first introduced. An analysis and/or reasoning as to the nomenclatural and taxonomic status of the name is a familiar component of a treatment;

4. Information about the etymology of the name; statements as to the correct, alternate or erroneous spellings; reference or citation to the literature where the name was introduced or changed;

5. Rank, composition and/or apomorphy of a taxon;

6. For species and subordinate taxa that have been placed in different genera, the author (with or without date) of the basionym of the name or the author (with or without date) of the combination or replacement name;

7. Lists of synonyms and/or chresonyms or taxon concepts, including analyses and/or reasoning as to the status or validity of each;

8. Citations of publications that include taxonomic and nomenclatural acts, including typifications;

9. Reference to the type species of a genus or to other type taxa;

10. References to type material, including current or previous location of type material, collection name or abbreviation thereof, specimen codes and status of type;

11. Reference to the registration number of the taxon or nomenclatural act (bounding information in mycology, voluntary in botany, zoology and paleontology);

12. Data about materials examined;

13. References to image(s) or other media with information about the taxon;

14. Information on overall distribution and ecology, perhaps with a map;

15. Known uses, common names and conservation status (including Red List status recommendation);

16. Description and/or circumscription of the taxon (features or traits together with the applicable values), diagnostic characters of a taxon, possibly with the means (such as a key) by which the taxon can be distinguished from relatives;

17. General information including, but not limited to: taxonomic history, morphology and anatomy, reproductive biology, ecology and habitat, biogeography, conservation status, systematic position, phylogenetic relationships of and within the taxon,

population-genetic diversity, structure and relationships within and between taxa and references to relevant literature;

18. Genomic information derived from an identifiable organism, an assemblage of organisms or eDNA, ranging from whole genome information to chromosome rearrangements, insertions and deletions, localised sequences, single nucleotide repeats (SSRs, microsatellites) or single nucleotide point mutations and more, as well as identifiers linking to such information in external repositories;

19. Photographs (or other image or series of images) by a person or persons using a recording device, such as a scanner or camera, whether or not associated with light- or electron-microscopes, using X-rays, acoustics, tomography, electromagnetic resonance or other electromagnetic sources, of whole organisms, groups, colonies, life stages especially from dorsal, lateral, anterior, posterior, apical or other widely used perspectives and designed to show overall aspect of organism*[1];

20. Photographs (or other image or series of images) by a person or persons using a recording device, such as a camera associated with light- or electron-microscopes, using X-rays, acoustics, tomography, electromagnetic resonance images or other electromagnetic sources) of parts of organisms, such as, but not limited to appendages, mouthparts, anatomical features, ultrastructural features, flowers, fruiting bodies, foliage, intra-organismic and inter-organismic connections, of compounds and analyses of compounds extracted from organisms that demonstrate the characteristics of an individual or taxon and/or allow comparison amongst individuals/taxa;

21. Photographs (or other images or series of images) of whole organisms, groups, colonies, life stages, parts of organisms made by camera or scanner or comparable devices using automated procedures;

22. Drawings of organisms or parts of organisms made by a person or persons to demonstrate the characteristics of an individual/taxon or to allow comparisons amongst taxa;

23. Graphical/diagrammatic representation (such as, but not limited to, scatter plots with or without trend lines, histograms or pie charts) of quantifiable features of one or more individuals or taxa for the purposes of showing the characteristics or allowing comparison of individuals or taxa and made by a person or persons.

Acknowledgements

As part of the CETAF e-Publishing Working group work on best practices in publishing, this paper was co-authored by members of three associations and their communities and reflects the discussions held during various meetings. Its rationale has been developed in a more detailed text to be published soon (Buschbom et al., in press). It was then approved

by their respective Executive boards. The text was improved by the comments of several colleagues that we would like to thank here: Andreas Kroh, Jiří Kvaček, Anahita Kazem, Michelle Price, Gergely Babocsay, Scott Rufolo, Martin Kalfatovic and David Iggulden. The authors would like to give a special recognition to Connie Rinaldo who was a founding group member, with Jutta Buschbom, Laurence Bénichou and Donat Agosti, contributing to the preparation of the joint-workshop at the TDWG meeting in 2022 before she passed away on the 27 October 2022.

Funding program

This initiative is supported by the BiCIKL project which receives funding from the European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Agosti D, Benichou 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 https://doi.org/10.3897/rio.8.e97374
- Anonymous (2014) The Bouchout Declaration for Open Biodiversity Knowledge Management. URL: https://www.bouchoutdeclaration.org/declaration.html
- Ball A (2014) How to License Research Data. A Digital Curation Centre and JISC Legal 'working level' guide. DCC How-to Guides. Edinburgh: Digital Curation Centre. URL: https://www.dcc.ac.uk/guidance/how-guides/license-research-data
- Benichou 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 https://doi.org/10.3897/rio.8.e94338
- 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 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
- Buschbom J, et al. (in press) Reuse of scientific data in scholarly publications. European Journal of Taxonomy.
- Carroll SR, Garba I, Figueroa-Rodríguez O, Holbrook J, Lovett R, Materechera S, Parsons M, Raseroka K, Rodriguez-Lonebear D, Rowe R, Sara R, Walker J, Anderson

J, Hudson M (2020) The CARE Principles for Indigenous Data Governance. Data Science Journal 19 https://doi.org/10.5334/dsj-2020-043

- 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 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 https://doi.org/10.3897/rio.3.e12502
- EOSC (2023) European Open Science Cloud Onboarding : How to Become an EOSC Provider - An Overview. Information Required about Providers and their Resources. URL: https://eosc-portal.eu/eosc-providers-hub/how-become-eosc-provider/how-become-eosc-provider-a-general-overview
- 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. Publications Office of the European Union. URL: https://data.europa.eu/doi/10.2777/891665
- European Parliament and Council (2019) 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. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32019L0790
- Gervais DJ (2019) (Re)structuring Copyright: A Comprehensive Path to International Copyright Reform. Edward Elgar Publishing, 384 pp. [ISBN 978 1 78990 214 3]
- Group on Earth Observations (2021) GEO Statement on Open Knowledge. URL: https://www.earthobservations.org/documents/geoweek2021/GEO-17-4.1_GEO%20Statement%20on%20Open%20Knowledge.pdf
- Kalkman S, Mostert M, Gerlinger C, van Delden JM, van Thiel GMW (2019) Responsible data sharing in international health research: a systematic review of principles and norms. BMC Medical Ethics 20 (1). https://doi.org/10.1186/s12910-019-0359-9
- OECD (2021) Recommendation of the Council on Enhancing Access to and Sharing of Data, OECD/LEGAL/0463. URL: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0463
- Patterson DJ, Egloff W, Agosti D, Eades D, Franz N, Hagedorn G, Rees JA, Remsen DP (2014) Scientific names of organisms: attribution, rights, and licensing. BMC Research Notes 7 (1). https://doi.org/10.1186/1756-0500-7-79
- 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, Benichou 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 https://doi.org/10.3897/rio.8.e81136
- Salwén H (2021) Research Ethical Norms, Guidance and the Internet. Science and Engineering Ethics 27 (6). https://doi.org/10.1007/s11948-021-00342-5
- SR 231.1 (1992) Federal Act on Copyright and Related Rights (Copyright Act, CopA) of 9 October 1992 (Status as of 1 July 2023). URL: https://www.fedlex.admin.ch/en/cc/internal-law/23#231

- UNESCO (2021) Recommendation on Open Science. URL: https://en.unesco.org/science-sustainable-future/open-science/recommendation
- U.S. Supreme Court (1991) Feist Publications, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340. URL: https://supreme.justia.com/cases/federal/us/499/340/
- Watanabe ME (2018) Digitizing Specimens—Legal Issues Abound. BioScience 68 (9): 728-728. https://doi.org/10.1093/biosci/biy086
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, 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 (1). https://doi.org/10.1038/sdata.2016.18

Endnotes

[*1] Numbers 19 to 21 are not applicable to some European countries that provide for a special protection for non-individual photographs (e.g. Austria, Denmark, Germany, Italy, Sweden and Switzerland).