

A deafening silence: a lack of data and reproducibility in published bioacoustics research?

Ed Baker^{‡,§}, Sarah Vincent[‡]

[‡] Natural History Museum, London, United Kingdom

[§] University of York, York, United Kingdom

Corresponding author: Ed Baker (edwbaker@gmail.com)

Academic editor: Therese Catanach

Abstract

A study of 100 papers from five journals that make use of bioacoustic recordings shows that only a minority (21%) deposit any of the recordings in a repository, supplementary materials section or a personal website. This lack of deposition hinders re-use of the raw data by other researchers, prevents the reproduction of a project's analyses and confirmation of its findings and impedes progress within the broader bioacoustics community. We make some recommendations for researchers interested in depositing their data.

Keywords

bioacoustics, open data, reproducibility, sound libraries, acoustic vouchers

Introduction

The importance of sharing the datasets used in biological research has been discussed recently by a number of authors, for example, in ecology (e.g. Poisot et al. 2013, Kenall et al. 2014), phylogenetics (e.g. Magee et al. 2014, Stoltzfus et al. 2012) and behaviour (e.g. Caetano and Aisenberg 2014). These authors list several benefits of sharing data, including the opportunity to create future collaborations and clarification of authorship. There can also be ethical considerations, including the use of public funds to generate these datasets. Another significant reason for sharing datasets underpinning research is to ensure that those findings are reproducible, a fact which advocates for open science have discussed well before the recent 'reproducibility crisis' in psychology (Open Science Collaboration 2015).

The datasets used in bioacoustic research vary in scale from a single short recording to continuous recordings at a site over several years. These recordings may be used to identify (e.g. Heller and Baker 2017) or describe (e.g. Dring 1983) species new to science and to facilitate long-duration biological surveys (e.g. Eichinski and Roe 2017). Additionally, recordings may be used to design automated surveys that do not themselves make recordings (e.g. Bennett et al. 2015).

Material and methods

The twenty most recently published articles (as of the end of 2017) covering bioacoustic topics were selected from each of the journals Bioacoustics, ZooKeys, ZooTaxa, Journal of Animal Behaviour and Marine Mammal Science. Primary research articles were identified using the search terms "acoustics" and "bioacoustics" on the journal's website; only articles making use of recorded sounds were selected. The journals chosen cover a subject-specific journal (Bioacoustics), a modern semantically enhanced (Penev et al. 2010) journal (ZooKeys), a taxonomic mega-journal (ZooTaxa), a journal with a broader zoological scope (Journal of Animal Behaviour) and a taxonomically focussed journal (Marine Mammal Science).

Both authors independently examined the papers for evidence that the underlying sound recordings were available in either physical or digital repositories. Two scoring systems were used to categorise the papers:

- The first identifies those papers that define a repository for sound recordings;
- The second differentiates between the type of deposition (repository, personal website, supplementary material) and whether the recordings deposited are the complete set or a representative sample.

Data resources

A CSV file of the papers analysed (including their DOIs) and scored values is available in Suppl. materials 1, 2, full Crossref metadata for the articles is available in Suppl. materials 3, 4 .

Results

Out of all the articles in the study, 79% did not deposit any sound recordings, 12% deposited a sample of the studied data and only 9% deposited the full acoustic dataset (Fig. 1). Of those articles depositing the full dataset, 78% were from the same journal (ZooKeys). ZooKeys was the only journal to achieve a deposition rate of over 50% when considering both complete and sample depositions (Fig. 2).

Where complete data were deposited, the majority used a repository (8%) or the supplementary materials of the journal article (7%). Only 2% used a personal website (Fig. 3).

Discussion

It can be seen from these results that only a minority (21%) of the published studies analysed deposit the sound recordings on which their results are based. In addition, even when authors claim deposition, there can still be problems. Issues identified in this study include not depositing recordings, broken URLs and providing the wrong URL for a repository (for a list see Suppl. material 1).

This lack of deposition is potentially problematic for the reproducibility of research and also hinders the re-use of recordings by other researchers. Given widespread lack of deposition described in this paper, below we provide some recommendations which researchers may use to improve the accessibility of their bioacoustic data.

Reproducibility

It may be argued for well-studied, easily identified taxa with relatively stable taxonomy, that there is no need to deposit recordings and that a well-documented methodology is sufficient to ensure reproducibility. The reasons why this does not extend to all taxa have been discussed previously in the context of voucher specimens for biodiversity and community ecology by Turney et al. (2015) and for phylogenetics by Pleijel et al. (2008). As many bioacoustic signals are unique to species, they can, in many cases, be considered to be surrogates for voucher specimens. As a minimum, we therefore recommend deposition of a sample of voucher sound recordings. Where it is feasible to collect voucher specimens and sound recordings, linking recordings to specimens in museum collections provides an even more robust identity for the organisms studied and allows the published study to remain relevant, even if the species studied is later found to be two or more species.

Making Data Available

Dedicated bioacoustic repositories often have the advantage of integrating with other components of the biodiversity informatics landscape, for example using Darwin Core (Wieczorek et al. 2012) to provide species locality data to the Global Biodiversity Informatics Facility (GBIF). This integration with external aggregators adds additional impact to the datasets that are shared; however it is not universally adopted by bioacoustic repositories. GBIF is perhaps the most prominent tool for searching for recordings of species across the datasets of multiple institutions and we strongly encourage institutions holding bioacoustic data to contribute.

Riede (2018) discusses potential depositories for the singing Orthoptera, the [Macaulay Library](#) and [xeno-canto](#) have large collections of birdsong. The [BioAcoustica](#) database (Baker et al. 2015b) accepts recordings of any species in addition to soundscape

recordings and makes data available to the Encyclopedia of Life and GBIF (Baker et al. 2015a). The choice of repository is down to individual researchers and the taxa they work with; however, considerations may include whether these repositories make the original sound files available online (instead of just MP3s which may not contain all relevant acoustic data), whether the files can be openly licensed (e.g. using Creative Commons) and whether off-site backups are provided to protect against data loss (e.g. Dena et al. 2018). Some repositories (e.g. the [Data Portal](#) at the Natural History Museum, London; Scott et al. 2019) provide DataCite DOIs for contributed datasets, allowing for individual datasets to be cited.

Many of the advances in large-scale ecoacoustics will rely on large datasets that are labelled suitably for machine learning algorithms. While these datasets are becoming available for well-studied groups of organisms (Morfi et al. 2019), expanding this to less well-studied taxa will require widespread data sharing to obtain comparable datasets in a reasonable timeframe.

Conclusions

More work needs to be done by the bioacoustics community to create an environment where the data underpinning research are made available, ideally using the FAIR Data Principles of being findable, accessible, interoperable and reusable. A discussion of the FAIR principles for scientific data is provided in Wilkinson et al. (2016). The loss of data due to natural or anthropogenic causes (e.g. Dena et al. 2018) is reduced by having an independent backup in an external location.

Recommendations

The recommendations below are based on issues we have identified in the research for this paper.

1. The use of a repository (either a formal repository or institutional data portal) is recommended for bioacoustic recordings to aid with the findability and accessibility components of the FAIR data principles. Consideration should be given to the long-term sustainability of the repository (e.g. institutional support), how that repository connects to the wider biodiversity informatics landscape and the formats in which it makes audio files available (some repositories only make lossily-compressed MP3 files available to end users even though they have WAV files available internally).
2. The repository where recordings will be deposited should be identified before a paper is submitted. Working with the repository will allow for URLs to individual recordings (or sets of recordings) to be included within the paper, instead of a generic reference to the repository that the end-user must then search or browse to uncover the recordings. Communication with the repository prior to publication should eliminate any issues of providing incorrect URLs in published works. Depositing recordings prior to submission will also prevent instances of papers

- claiming submission to a repository, but the authors then forget to make such depositions.
3. Repositories may allow for an embargo on the public release of recordings until a paper is published and/or for a time-limited period subsequently.
 4. The use of the Supplementary Materials section of journals is not recommended for audio deposition, as access may be limited to journal subscribers and they are not at present discoverable via aggregators such as GBIF.
 5. Repositories should have a mechanism to prevent link-rot from changing URLs, ensuring that cited URLs are always resolvable.
 6. Use of an open licence such as Creative Commons allows data to be re-used easily by other researchers.

Conflicts of interest

References

- Baker E, Price B, Rycroft S, Villet M (2015a) Global Cicada Sound Collection I: Recordings from South Africa and Malawi by B. W. Price & M. H. Villet and harvesting of BioAcoustica data by GBIF. *Biodiversity Data Journal* 3 (5792): e5792. <https://doi.org/10.3897/BDJ.3.e5792>
- Baker E, Price BW, Rycroft SD, Hill J, Smith VS (2015b) BioAcoustica: a free and open repository and analysis platform for bioacoustics. Database 2015: bav054. <https://doi.org/10.1093/database/bav054>
- Bennett W, Chesmore D, Baker E (2015) Speckled bush cricket data logger - Project Report. Figshare <https://doi.org/10.6084/M9.FIGSHARE.1430094.V1>
- Caetano D, Aisenberg A (2014) Forgotten treasures: the fate of data in animal behaviour studies. *Animal Behaviour* 98: 1-5. <https://doi.org/10.1016/j.anbehav.2014.09.025>
- Dena S, Rebouças R, Augusto-Alves G, Toledo LF (2018) Lessons from recordings lost in Brazil fire: deposit and back up. *Nature* 563 (7732): 473-473. <https://doi.org/10.1038/d41586-018-07456-0>
- Dring J (1983) Some new frogs from Sarawak. *Amphibia-Reptilia* 4 (2): 103-115. <https://doi.org/10.1163/156853883x00021>
- Eichinski P, Roe P (2017) Clustering and visualization of long-duration audio recordings for rapid exploration avian surveys. 2017 IEEE 13th International Conference on e-Science (e-Science) <https://doi.org/10.1109/eScience.2017.29>
- Heller K-, Baker E (2017) From an old sound recording to a new species in the genus *Horatosphaga* (Orthoptera: Tettigoniodea: Phaneropterinae: Acrometopini). *Zootaxa* 4323 (3): 430. <https://doi.org/10.11646/zootaxa.4323.3.10>
- Kenall A, Harold S, Foote C (2014) An open future for ecological and evolutionary data? *BMC Evolutionary Biology* 14 (1): 66. <https://doi.org/10.1186/1471-2148-14-66>
- Magee A, May M, Moore B (2014) The dawn of open access to phylogenetic data. *PLoS ONE* 9 (10): e110268. <https://doi.org/10.1371/journal.pone.0110268>

- Morfi V, Bas Y, Pamula H, Glotin H, Stowell D (2019) NIPS4Bplus: a richly annotated birdsong audio dataset. *PeerJ Computer Science* 5 <https://doi.org/10.7717/peerj-cs.223>
- Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science* 349 (6251): aac4716. <https://doi.org/10.1126/science.aac4716>
- Penev L, Agosti D, Georgiev T, Catapano T, Miller J, Blagoderov V, Roberts D, Smith V, Brake I, Ryrccroft S, Scott B, Johnson N, Morris R, Sautter G, Chavan V, Robertson T, Remsen D, Stoev P, Parr C, Knapp S, Kress WJ, Thompson F, Erwin T (2010) Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. *ZooKeys* 50: 1-16. <https://doi.org/10.3897/zookeys.50.538>
- Pleijel F, Jondelius U, Norlinder E, Nygren A, Oxelman B, Schander C, Sundberg P, Thollesson M (2008) Phylogenies without roots? A plea for the use of vouchers in molecular phylogenetic studies. *Molecular Phylogenetics and Evolution* 48 (1): 369-371. <https://doi.org/10.1016/j.ympev.2008.03.024>
- Poisot T, Mounce R, Gravel D (2013) Moving toward a sustainable ecological science: don't let data go to waste! *Ideas in Ecology and Evolution* 6 (2). <https://doi.org/10.4033/iee.2013.6b.14.f>
- Riede K (2018) Acoustic profiling of Orthoptera: present state and future needs. *Journal of Orthoptera Research* 27 (2): 203-215. <https://doi.org/10.3897/jor.27.23700>
- Scott B, Baker E, Woodburn M, Vincent S, Hardy H, Smith VS (2019) The Natural History Museum Data Portal. Database 2019: baz038. <https://doi.org/10.1093/database/baz038>
- Stoltzfus A, O'Meara B, Whitacre J, Mounce R, Gillespie EL, Kumar S, Rosauer DF, Vos RA (2012) Sharing and re-use of phylogenetic trees (and associated data) to facilitate synthesis. *BMC Research Notes* 5 (1): 574. <https://doi.org/10.1186/1756-0500-5-574>
- Turney S, Cameron E, Cloutier C, Buddle C (2015) Non-repeatable science: assessing the frequency of voucher specimen deposition reveals that most arthropod research cannot be verified. *PeerJ* 3 <https://doi.org/10.7717/peerj.1168>
- Wiczorek J, Bloom D, Guralnick R, Blum S, Doering M, Giovanni R, Robertson T, Vieglaisset D (2012) Darwin Core: An evolving community-developed biodiversity data standard. *PLoS ONE* 7: e29715. <https://doi.org/10.1371/journal.pone.0029715>
- Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, 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>

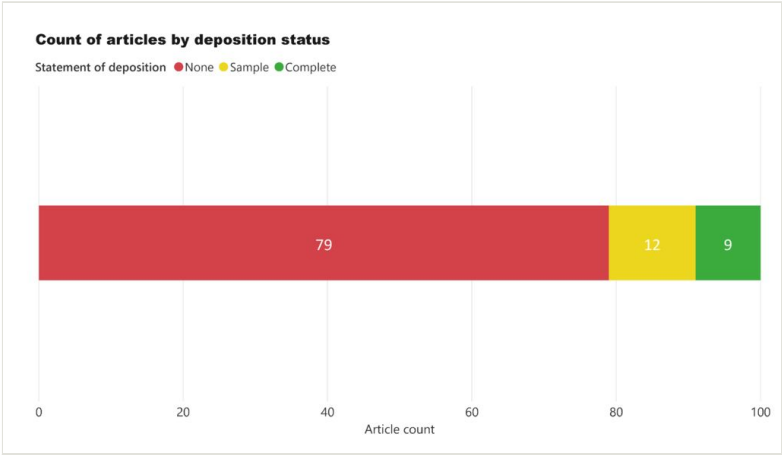


Figure 1.
Deposition status of all articles in the study.

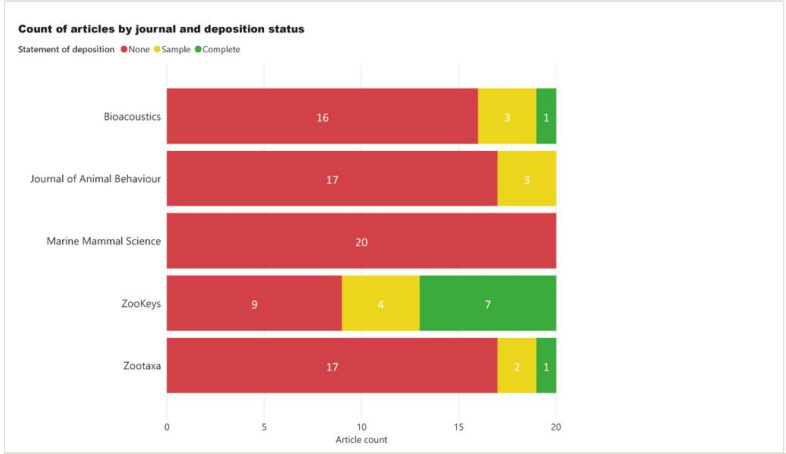


Figure 2.
Breakdown of article deposition status by journal.

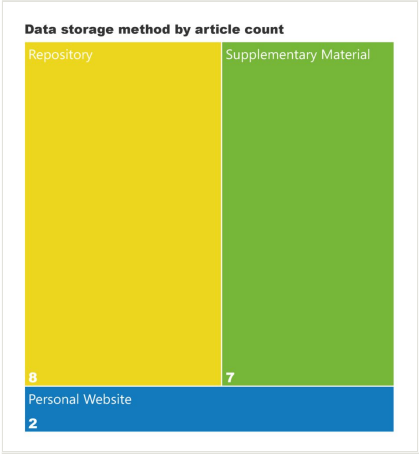


Figure 3.
Deposition method for papers that deposit data.

Supplementary materials

Suppl. material 1: Scoring of Articles (CSV)

Authors: Ed Baker; Sarah Vincent

Data type: references

Brief description: The scoring of the articles used in this study.

[Download file](#) (20.96 kb)

Suppl. material 2: Scoring of Articles (Excel)

Authors: Ed Baker; Sarah Vincent

Data type: references

Brief description: The scoring of articles used in this study.

[Download file](#) (20.92 kb)

Suppl. material 3: Crossref Metadata for papers analysed (CSV)

Authors: Sarah Vincent

Data type: references

Brief description: Full metadata for the articles analysed.

[Download file](#) (29.88 kb)

Suppl. material 4: Crossref Metadata for papers analysed (Excel)

Authors: Ed Baker; Sarah Vincent

Data type: references

Brief description: Full metadata for the articles analysed.

[Download file](#) (27.79 kb)