Essay

Forgotten treasures: the fate of data in animal behavior studies

Daniel S. Caetano[1,*] and Anita Aisenberg[2]

[1] Department of Biological Sciences, University of Idaho, Moscow, ID 83844, USA. E-mail: caetanods1@gmail.com.

* Corresponding author.

[2] Laboratorio de Etología, Ecología y Evolución, Instituto de Investigaciones Biológicas Clemente Estable, Avenida Italia 3318, Postal Code 11600, Montevideo, Uruguay. E-mail: anita.aisenberg@gmail.com

Corresponding author e-mail: caetanods1@gmail.com

Keywords: database; data reuse; data sharing; science policy; museum collections

Abstract

Published discussions on data stewardship often focus on standardized datasets whose reuse patterns are known. Improvements in stewardship of animal behavior data are virtually absent and lag behind other disciplines such as molecular biology and systematics. In this essay, we discuss best practices of three key aspects related to the collection and archival of behavioral data: data supporting published results; data collected from field observations; and the potential of museum specimens as source of data to animal behavior and ecology. To quantify how much data is shared in publications we reviewed selected journals in animal behavior and behavioral ecology. We found that only an extremely small proportion of the articles published in 2013 made even part of their data available. We discuss about the benefits of making data available, review resources available for data archiving and provide practical guidance for ethologists. We discuss and provide examples of the amount of ethological and ecological data that can be recorded during field observations. To investigate the potential of museum specimens as source of data, we surveyed researchers working in areas related to ecology, animal behavior, and systematics. Both ethologists and systematists agreed that natural history information stored in collections would be a valuable source of data. We make recommendations to enhance data collection and stewardship from the point of view of researchers in animal behavior sciences, considering the special characteristics of the discipline and the type of data that is often produced. We suggest that there is a large amount of crucial data about natural history, ecology and behavior that investigators could glean from collections. Although it is difficult to appreciate the relevance of data for future studies at the time of publication, such data may inspire fruitful opportunities that we cannot afford to lose.

Data collection is a fundamental step of research and often the most expensive with respect to both time and money, yet the fate of data after publication is often neglected (Heidorn, 2008). Despite willingness to populate digital repositories with data, the fact that many researchers have reservations about doing so (Tenopir et al., 2011; Wolkovich et al., 2012) seems to explain the general lack of data available for publications in the biological sciences (Hartter et al. 2013; Zamir, 2013 – but see Wallis et al., 2013). There are a few exceptions, such as molecular journals that require sequence data archival in publicly accessible repositories among other examples such as the journals Evolution and the Public Library of Science (PLoS). The main reasons why authors opt to avoid storing data in digital repositories are related to concerns about lack of time and appropriate tools to prepare and upload datasets, the potential for data misuse (Whitlock, 2011), and lack of personal benefits (Arzberger et al., 2004). In this essay, we surveyed and evaluated how much data from animal behavior is made available in public repositories. We report that there are tools available for data sharing, which are easy to use and of low cost, and provide practical guidelines to data management specific to animal behavioral data. We also point out that museum collections are an underexploited source of animal behavior data whose access is facilitated by recent investments in the digitalization of collections worldwide.

Internet has facilitated advances in scientific communication, which scale from email lists and social media groups (twitter, blogs, reddit, etc.) to articles published on-line prior to (e.g., bioRxiv and PeerJ pre-prints) or immediately following the peer-review process. One would expect that broadband communication would not only enhance access to articles but also to the data that support them (Arzberger et al., 2004; Costello, 2009). However, absence of a data sharing culture among researchers in the biological sciences contradicts this expectation (Wolkovich et al., 2012). Access to digital storage space and limited capabilities to exchange datasets were technical issues in the 80s and early 90s. However, initiatives to promote open science and reproducibility over the last decade have led to increased availability of suitable resources to help manage, archive, and share data. The frequency of data archiving in publications has been estimated for ecological (Hampton et

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.396v1 | CC-BY 4.0 Open Access | received: 26 May 2014, published: 26 May 2014

al., 2013; Vines et al., 2013), evolutionary (Drew et al., 2013) and health sciences (Piwowar, 2011) publications. However, to the best of our knowledge, no one has surveyed the frequency of data from animal behavior publications made available in digital repositories or supplementary materials.

Quantifying animal behavior data availability

We randomly selected and reviewed one third of the articles published during 2013 in Animal Behaviour (AB – 103 out of 308 articles) and Behavioral Ecology (BE – 54 out of 161 articles). We chose to sample from those journals because we recognize them to be among the most influential journals in animal behavior sciences. We searched for database indications (hyperlinks and/or references) in the methods, results and acknowledgements sections of each publication (see Table S1 in DOI: 10.6084/m9.figshare.1003857). We recorded whether the article reported summary statistics (e.g., proportions, mean/median and standard error) in the text or in tables, and if at least part of the raw data was made available in tables, supplementary material or stored in a digital repository (Figure 1). Summary statistics are often reported in figures (e.g., histograms and box plots), but we did not include figures in our survey because it is impracticable and often impossible to recover the original values used to generate them (but see DataThief III – Tummers, 2006). Figures are an efficient media to show contrasts and relationships among results; however, they are not a good means to report data.

Only a small proportion of the analyzed articles from Animal Behaviour (13%) and Behavioral Ecology (7%) made at least some portion of their data available. Although our sample is restricted to one year, our results are similar to a survey of environmental biology publications over a five years period that reported only 8% of articles sampled made their data available (excluding sequence data) (Hampton et al., 2013). The majority of publications that we surveyed reported summary statistics (AB: 63%; BE: 68.5%). Of these, a minority reported summary statistics in tables (AB: 41.5%, n = 27 out of 65; BE: 27%, n = 10 out of 37). A majority of reports within the

text is expected for Animal Behaviour publications, since the journal asks for authors to be sparing in the use of tables. This expectation does not apply for Behavioral Ecology papers, yet those showed relatively less frequency of results reported in tables. The use of tables to report summary statistics is advised, because they can facilitate collating results from different studies for data mining and meta-analyses (but see Noorden, 2014). Summary statistics, parameter estimates, results from tests of significance and effect sizes (Nakagawa & Cuthill, 2007) are the main information needed to understand the findings and conclusions of a scientific publication and to perform meta-analyses. However, they cannot be considered as data, since they are a product of interpretations that authors have assigned to their observations and do not allow for reproduction of the findings. For example, it is difficult to discern the degree of individual-level variation contained within behavioral categories when only the results of the categorization are provided. If the raw data are available, researchers can re-evaluate those categories. Furthermore, since there are a wide range of ways to define and evaluate individual characteristics (e.g., body condition – Moya-Laraño et al., 2008), it is crucial to have access to details of any assumptions that the authors made in order to adapt results from one study system to another. Lack of data availability can make evaluations challenging and create barriers to scientific communication.

The cost of losing data

Failure to store data from animal behavior studies comes at a big cost. The majority of studies results from the observation of a cohort of individuals in a specific point in time and space (Heidorn, 2008; Wolkovich et al., 2012). Behavioral plasticity, geographic variation and environmental fluctuations make the reproducibility of such studies challenging (see related discussion in Bissell, 2013). As a result, specific behavioral data not made available in repositories are likely going to be lost. There are different ways to ensure that these data are not lost. For instance, journals can require that authors share data (Whitlock et al., 2010; Alsheikh-Ali et al., 2011). However, this policy is uncommon in animal behavioral journals. None of the journals

classified under the behavioral sciences category in the Journal of Citation Reports database (ISI Web of Science) require data archiving following publication. Although all journals accept supplementary data from a range of media formats (e.g., sound, video and photos), less than half (34%, n = 49) explicitly encourage authors to make the data available in digital repositories and none require data archiving prior to publication (see Table S2 in DOI: 10.6084/m9.figshare.1003857). Furthermore, despite journal policies, it is not clear if authors comply. Savage and Vickers (2009) asked authors of articles published in the PLoS journals to share their datasets and, contrary to the journal, an impressive portion of authors refused to do so (see also Wicherts et al., 2006; Alsheikh-Ali et al., 2011). In 2014, perhaps as a reaction to those issues, PLoS journals started to require authors to deposit their data in publicly assessable repositories prior to publication (Bloom et al., 2014). Additionally, funding agencies also commonly require that authors share data of publications (Costello, 2009). On the other hand, if data management plans are based solely on journal or funding requirements, a significant amount of data may be lost (Savage & Vickers, 2009). Proper data stewardship and sharing is good scientific practice and should therefore not be viewed simply as a mandatory requirement to fulfill (Costello, 2009; Piwowar & Vision, 2013).

One common alternative to make data available after publication is adopted by authors who share datasets upon request from the scientific community. Some behavioral sciences journals share this view and require that authors provide datasets when requested, yet remain agnostic to whether data should be deposited in repositories (see Table S2 in DOI: 10.6084/m9.figshare.1003857). Vines and collaborators (2013) showed that datasets not stored in a repository rapidly tend to be lost over time – 80% of the data is lost within 20 years. Sharing data prevents this loss, since datasets available to public reuse are more likely to survive in the long term (Gibney, 2013). On the other hand, data stored on private hard-drives or local repositories are often lost due to disuse (Heidorn, 2008; Wolkovich et al., 2012). Researchers, funding agencies, and institutions are more prone to be concerned with large datasets resulting from collaborations and/or associated with long-term

projects (Heidom, 2008). However, the majority of published studies, especially in the animal behavior sciences, produce smaller datasets due to characteristics of the study system or experimental design. The heterogeneity of datasets and difficulties with reproducibility are the main reasons why losing data from animal behavior studies is of particular concern. Each dataset represents a spatial, temporal or population replicate of importance to future studies, but scientists often fail to recognize such potential at the time of publication (Wolkovich et al., 2012).

Altruistic or selfish behavior? Neither one, nor the other

The archival of data in digital repositories and metadata management is the responsibility of the authors (see discussion in Roche et al., 2014). At first inspection, this practice seems to be an altruistic behavior, beneficial to the community with no individual return. Individual benefit is among the main concerns of researchers when questioned about data sharing (Costello, 2009; Wolkovich et al., 2012). Contrary to this perception, there are benefits associated with data sharing both at the individual and community level (Craig et al., 2007; Costello, 2009; Piwowar & Vision, 2013).

Articles with data publicly available receive more citations. Piwowar and Vision (2013) showed that articles which data are available receive an overall 9% increase in the number of citations after correcting for confounding variables. In addition, data publicly available increase visibility of articles to Internet searches (Piwowar et al., 2007) and is likely to be indexed by search databases such as DataCite (http://datacite.org) and Data Citation Index (ISI Web of Science). Data availability provides transparency to publications. Peers who are not able to access the data supporting a publication need to "trust" the authors and the review process, which is also closed in most of the journals. Lack of transparency makes publications vulnerable to acts of scientific misconduct, what can damage the credibility of individuals, institutes and funding agencies (Couzin, 2006; Costello, 2009). At the individual level, transparency may increase citations due to more confidence of the peers in the results reported by the authors (Costello, 2009; Piwowar &

Vision, 2013). Starting in 2013, the National Science Foundation (NSF) asks researchers applying for grants to list their products rather than their publications. This means that not only research articles, but other types of products, such as datasets, are recognized by NSF as scientific production (Piwowar, 2013). Datasets in digital repositories can be identified using Digital Object Identifiers (DOI – see Table 1) that allow products to be cited in journal articles and have their impact in social media tracked through services such as ImpactStory (http://impactstory.org) and Altmetric (Priem et al., 2010). Costello and collaborators (2013a) go further and state that datasets should be published in specialized journals or dedicated sections after peer-review under the same standards of a research manuscript. Independent of introducing new metrics or publishing datasets in journals, the push for recognizing data as research production is strong (Piwowar et al., 2011; Wolkovich et al., 2012; Costello et al., 2013a; Drew et al., 2013; Piwowar, 2013; Vines et al., 2013). Both scientific community and individual researchers may be rewarded by the establishment of a data sharing culture in animal behavior sciences. Use of datasets without proper citation would likely be discouraged, researchers would receive recognition for reused products and articles would become more transparent and reproducible (Wolkovich et al., 2012; Costello et al., 2013a; Piwowar, 2013).

Guidelines to animal behavior data sharing and archiving

Animal behavior studies often record data in a myriad of media formats including images, videos, and audio recordings. Although the unique characteristics of such media make archival in a standardized format (such as molecular sequences in GenBank) difficult (Benson et al., 2014), there are several digital repositories capable of storing such heterogeneous datasets (Table 1). Most of the repositories do not charge for data deposition. Those repositories also offer a limited amount of private storage space and unlimited storage space for released datasets. Scientists may use private storage space to archive data while conducting research. Beyond serving as a reliable back-up, this practice improves data management efficiency as data are uploaded and organized at the moment of

publication. Among repositories listed in Table 1, Dryad releases datasets in the public domain under a Creative Commons Zero license (CC0) and fig**share** uses CC0 for data and Creative Commons Attribution (CC-BY) for media files. The public domain license (CC0) waives all legal requirements to attribution of rights to the authors whereas the attribution license (CC-BY) requires citation of the original authors, reproduction of copyright notices present in the work, and acknowledgement of modifications made to the original work. Both Zenodo and Maculay Library offer flexible license options. Some practical guidelines relevant to researchers preparing datasets to share via digital repositories include:

- **Record all metadata.** Metadata are information that describe data collection, defines categorizations, specifies data structure, and contains everything needed for another researcher to understand the data. A dataset with insufficient metadata can be impossible to reuse. Metadata also help in preventing errors in data collection of studies attempting to reproduce an experiment or incorporate more data to an available dataset. Well-constructed metadata make it possible for another researcher to thoroughly understand how the data were collected and, as a result, may facilitate collaboration in future research projects.

- **Implement extensive use of repositories.** It is often the case that authors publish only a portion of the data generated by research projects within their articles. As an alternative, authors can ensure that any data not directly related to the published results are available in repositories and assigned to DOIs. These data can be cited as soon as they are made available, foster collaborations, and bring visibility to young scientists. Availability of additional data is of special relevance to animal behavior studies in which observations of "rare" behaviors are often not reported. New findings may potentially be uncovered by comparative studies of rare behaviors (Peretti, 2013). However, those initiatives are made impossible due to the lack of accessibility to the data.

- **Deposit supplementary information in repositories.** Many journals provide the option for authors to include supplementary data that are then made available in the digital version of the publication. However, repositories increase data discoverability, are more reliable for long-term storage, and make it possible to make datasets available under open-access even when authors transfer copyright ownership to publishers.

- **Keep an eye on copyright licenses.** Datasets and media in publicly accessible repositories are usually shared under the Creative Commons attribution licenses (usually CC-BY) or in the public domain (CC0). The Creative Commons organization (https://creativecommons.org/) has extensive information on the different versions of the attribution family of licenses. The academic model of recognition by proper citation of authorship does not depend on legal requirements associated with copyright licenses and more restrictive licenses can create unwanted barriers to data reuse. Poisot and colleagues (2013) provide an interesting discussion on the application of licenses to shared datasets.

- **Embargo periods to release data.** One common concern of sharing data is that a third party could publish findings based on the dataset before the original authors. However, some data repositories have optional embargo periods that would prevent the release of datasets for a specified time period after the publication of the first article based on the dataset. One year seems a reasonable period to assure the "right of first use" to authors, but it is possible to request longer periods through repositories and/or journals (see discussion in Roche et al., 2014). Specific surveys are needed to estimate reasonable embargo periods for animal behavior datasets.

- **Establish a data management plan.** Data management plans help to keep datasets organized during data collection and provide a means to prevent data loss or recording errors and increase efficiency by diminishing data stewardship efforts posterior to data collection. Procedures to collect and record metadata, assure compatibility of file formats, and chose reuse license to release the data are examples of decisions that can be planned ahead of time. There are on-line tools available such as the DMPTool (https://dmp.cdlib.org/) and DMPonline (https://dmponline.dcc.ac.uk/) designed to help researchers create data management plans in the format required by different funding agencies.

Although there are plenty of resources for archiving data, it is important to stress that journals should implement policies related to data reuse in order to make researchers more comfortable with sharing their datasets. Several considerations include improved flexibility of embargo periods and the establishment of ethical standards to publishing results based on shared datasets (Roche et al., 2014).

Animal behavior and ecology data in museums

Although most data in the animal behavior sciences are gathered during standardized experiments or planned observations in the field, scientists also record incidental behavioral and ecological information when collecting specimens. Expeditions often visit scarcely sampled or unvisited sites and the material obtained can take years to process. Due to financial limitations or habitat loss and alteration, it is often not possible to resample sites (Fontaine et al., 2012). In such cases, only a relatively small number of individuals collected from a single study site are available for taxonomical studies and new species are described without prior knowledge of population level variation or range of geographic distribution (Fontaine et al., 2012; Costello et al., 2013b). As a result, data gathered during collection serve as the only available source of information. Those limitations to sample and observe specimens in the field are among the main reasons for our little

knowledge about the biology or ecology of a significant proportion of the described biodiversity (Costello et al., 2013b; Losos et al., 2013). Collectors usually keep detailed records in personal field books (Beidleman, 2004), but these data are often not linked to the specimens through labels or database access codes and may therefore be easily lost or inaccessible.

The importance of museum collections for biological research is indisputable (Winston, 2007), but collections are traditionally known to only provide taxonomical, morphological, geographical, and, more recently, molecular data. To our knowledge, the degree to which animal behaviorists and ecologists use data from collections is not known. We therefore surveyed scientists working in fields related to ecology, animal behavior, taxonomy, and systematics to investigate this question and to determine their perceptions of the type and amount of ecological and ethological information recorded and maintained in museum databases. For simplicity, we will use the terms 'ecologists' and 'taxonomists' to refer to scientists who self-reported as mainly working with ecology and animal behavior or taxonomy and systematics, respectively. We invited 381 researchers representing a range of institutions (e.g., universities, museums, and research facilities), primarily from the United States and Uruguay, to complete a survey. Professors, curators, and post-doctoral associates comprised the majority of participants (62%), followed by graduate students (32%), and undergraduate students (5%) (see Table S3 for survey questions and Table S4 for detailed results in DOI: 10.6084/m9.figshare.1003857).

Although survey participants agreed that all data should be recorded, they differed with respect to the type of data they deemed most essential for scientists to record. Compared to taxonomists, ecologists considered ecological observations (49% of 150 taxonomists vs. 69% of 152 ecologists), time of the day (28% vs. 44%), habitat descriptions (69% vs. 86%), and behavioral observations (28% vs. 41%) of greatest importance. Taxonomists argued that researchers should record data in field notebooks or databases whereas ecologists prefer to record and store information with the specimen (e.g., collection labels and/or databases with direct link to specimens). Direct linkage of ecological and behavioral data to specimens is a more reliable archiving procedure. Data in field

notebooks are prone to be lost and changes in taxonomical nomenclature or error in species identification can make information useless if the specimens cannot be re-evaluated. The set of questions pursued by taxonomists and ecologists usually are fairly distinct. Taxonomists recognize sampling locality, date of collection and name of collectors as the essential data required for taxonomical and systematic studies. In contrast, they are more prone to treat ecological observations as additional data only. On the other hand, ecological and behavioral data are paramount for ecologists. Despite the discrepancy between those disciplines, the study of biodiversity, its patterns and processes are common interests that may connect them and foster data interchange. We asked how often museum collections are used by ecologists and the frequency in which they rely on collaborations with taxonomists. More than half of the ecologists that completed the survey (63%) rely on zoological collections as a source of information for the majority (23%) or at least a portion (40%) of their research. Others do not rely on these data, yet they often work in collaboration with taxonomists (59%). Communication among disciplines is necessary in order to organize initiatives to enhance the amount of ecological and ethological information available for a significant portion of the known biodiversity.

Both the majority of taxonomists (91%) and ecologists (88%) agreed that ecological and behavioral information stored in zoological collections would serve as a good source of data for future studies. We received a total of 242 voluntary comments in response to this question (127 from systematists and 115 from ecologists). The majority of the systematists argued that inclusion of additional data would be valuable to museum collections, for it could provide justification for funding requests and would facilitate comparative studies. However, some participants suggested that data from independent observations would not be useful if not linked to published research. The opinions of ecologists and ethologists are divided: half argued that any piece of information is of great value, since encounters with rare species are uncommon and a surprisingly large number of species are only known as preserved specimens. Therefore, anecdotal remarks may provide the only source of information and the preservation of this data is crucial for future studies. The other half of

ecologists held an opinion contrary to this view; they argued that the cost in time and money to archive these data supersedes the potential benefits. The heterogeneity in the answers among ecologists may be related to the specific research interests of the scientists interviewed. Information collected haphazardly may be of little use for understanding consistent patterns across populations or species but can be valuable for within species variation studies, especially in the case of observations of rare occurrences.

Only a small fraction of the estimated total biodiversity is known and new species are often discovered during collecting expeditions. The opportunity to record information associated with the occasional observation of a specimen can be a rare event for particular taxa and could help improve our understanding of the real magnitude of Earth's biological diversity. There is an impressive amount of ethological and ecological data one can gather by encountering living specimens in the field (e.g., Caetano & Machado, 2013; Ramírez et al., 2013). Such data are crucial for comparative studies (e.g., Caetano & Machado, 2013), stimulate future research, or point to potential new study systems (e.g., Machado et al., 2004). In our view, the effort to archive and make those data available is justifiable. We have no means to estimate the value of the data lost when scientists fail to record detailed observations.

Zoological collections can provide invaluable animal behavior data

Although field observations are the most obvious source of animal behavior and ecology data, museum specimens can also provide information that is easily overlooked. One emblematic example is the exhaustive examination and description of female and male genitalia usually present in species descriptions and phylogenetic studies. Genitalia can provide details that are essential to understand the sexual strategies and the degree of sexual conflict in a species (Eberhard, 1985; Arnqvist, 2005). One example in arthropods is the occurrence of glandular mating plugs that cover total or partially the female genitalia (Uhl et al., 2010). Those plugs are frequently removed from preserved specimens in collections in order to allow a complete view of the genitalia external

anatomy and most of the times the information regarding them is lost. Descriptions of characteristics such as size, color, aspect, and frequency of occurrence of mating plugs can provide useful data that, combined with behavioral studies, will help determine the strength of male-male competition and female mating frequencies. Another example is the occurrence of male palpal breakage on the female epigynum in some spider species. In several species belonging to the families Araneidae and Theridiidae males break part of their palpal organ, the embolus, after insertion in the female insemination duct. This structure remains at the entrance of the duct and can function as mating plugs (Schneider & Andrade, 2011). Once again, both the occurrence and number of broken embolus in female genitalia could be easily quantified if these structures were not lost during morphological analysis, but preserved alongside the specimen. Also, the application of the clove oil frequently used to clarify genitalia of arthropods (Levi, 2004), can provide useful data as the occurrence of full or empty spermathecae, what informs us about female mating status.

Basic biological, ethological, and ecological data can be useful for a broad series of investigations; (a) Sexual selection studies: these data are essential for determining operational sex ratios, male-male competition, female choice, and levels of sexual conflict; (b) Conservation: information about the natural history and phenology can help identify potentially threatened species or ecosystems, contributing to motivate further studies to determine conservation status and plans for mitigation; (c) Comparative analyses: tests of evolutionary hypotheses require phylogenetic trees and phenotypic datasets. Those studies are improving our knowledge about patterns of diversification and their relationship with lineages' traits (e.g., Blackledge et al., 2009; Lapiedra et al., 2013); (d) Biological interactions: characteristics of the male and female genitalia can have valuable information on intra-specific interactions, whereas evidence of inter-specific interactions, such as epizoism, phoresy, and parasitism could be recorded during external and internal morphology analysis (e.g., Lücking et al., 2010; Penney et al., 2012); (e) Niche modeling: the incorporation of micro-habitat description and habitat use patterns in modeling algorithms make it possible to generate better projections of species distributions (see examples in Raxworthy et al.,

2007). Therefore, both recording additional information in the moment of the specimen collection and preserving the information available in museum specimens provide an important resource for future studies.

Future recommendations

One of the main reasons why data are not shared is the effort needed to organize datasets and manage descriptive metadata. Tools to facilitate data stewardship and mitigate the effort needed to submit datasets are imperative for the establishment of a data sharing culture in animal behavior sciences. Mobile devices such as smartphones and tablets have the potential to become important tools for data sampling. Camera and microphones are present in most mobile devices and development of applications specific to data collection is feasible (Powell, 2012). Direct connection with databases could allow researchers to update data from the field alongside basic information such as date, time, geo-reference and weather conditions. Researchers could store and edit data of ongoing projects in private (or public) databases (Table 1; also see the 'dat' project: http://dat-data.com/). Although Internet coverage in sampling sites is rare, initiatives such as Google's Loon project (http://www.google.com/loon/) point to networking advances that could amplify drastically the availability of connection in otherwise isolated locations.

The use of software to store data in databases through direct links would make data organization and metadata management automatic, thus requiring no further effort from researchers. This same work-flow could be easily implemented to link ecological and animal behavior information to museum databases. Data identified by unique codes provided by the collection's database program could be linked to the specimen using the same code printed on the specimen label (e.g., barcode). Implementation of specialized tools to merge data sampling, metadata collection and archiving into a single step should reduce time invested in organizing data after collection to a minimum. Since there are clear benefits to data sharing, we hope that mitigation of the effort related to data stewardship combined with recognition of datasets as scientific production

and the establishment of clear data reuse policies will help encourage data sharing among animal behaviorists.

Conclusions

Data sharing is not an altruistic behavior without benefits for the individual "productivity fitness". Datasets can be cited, funding agencies such as NSF recognize shared datasets as scientific products and articles that have data available are more likely to receive citations. The frequency that data is made available in animal behavior sciences is extremely low, as a result most of the data supporting publications are likely to be quickly lost. Museum specimens are an impressive source of ecological and animal behavior information that seems to be underexploited. New tools for data management and deposition on digital repositories are made available in response to the reproducibility movement in virtually all scientific disciplines. Data sharing is good scientific practice and should be more encouraged among animal behaviorists.

Acknowledgments

We thank G. Machado, A.B. Kury, T.H. Kawamoto, and F. Machado for fruitful discussions on the topic. B.A. Buzatto, L.E. Costa-Schmidt, A. Espíndola, D. Jochimsen, B.S. Medeiros, D.J. Machado, M.W. Pennell, A.V. Peretti, E.A. Santos, M. Simó, and J.C. Uyeda for comments on early versions of the manuscript. DSC is supported by fellowship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES – 1093/12-6). AA is supported by Programa Desarrollo de Ciencias Básicos (PEDECIBA), UdelaR, and Sistema Nacional de Investigadores (SNI), Agencia Nacional de Investigación e Innovación (ANII).

References

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H. & Ioannidis, J. P. A. (2011). Public availability of published research data in high-impact journals. *PLoS ONE*, 6, e24357.

Arnqvist, G., & Rowe, L. (2005). *Sexual conflict.* Princeton, NJ, U.S.A.: Princeton University Press.

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., 'et al.' (2004). An international framework to promote access to data. *Science*, 303, 1777–1778.

Beidleman, R. (2004). More than specimens in natural history museums. *BioScience*, 54, 6-7.

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. (2014). GenBank. *Nucleic acids research*, 42, D32–37.

Bissell, M. (2013). Reproducibility: The risks of the replication drive. *Nature*, 503, 333–334.

Blackledge, T. A., Scharff, N., Coddington, J. A., Szüts, T., Wenzel, J. W., Hayashi, C. Y., 'et al.' (2009). Reconstructing web evolution and spider diversification in the molecular era. *Proceedings of the National Academy of Sciences*, 106, 5229-5234.

Bloom, T., Ganley, E. & Winker, M. (2014). Data access for the open access literature: PLOS's data policy. *PLoS Biol*, 12, e1001797.

Caetano, D. S., & Aisenberg, A. (2014). Data for: "Forgotten treasures: the fate of data in animal behavior studies". *figshare*, doi: 10.6084/m9.figshare.1003857

Caetano, D. S., & Machado, G. (2013). The ecological tale of Gonyleptidae (Arachnida, Opiliones) evolution: phylogeny of a Neotropical lineage of armored harvestmen using ecological, behavioral, and chemical characters. *Cladistics*, 26, 589–609.

Costello, M. J. (2009). Motivating online publication of data. *BioScience*, 59, 418–427.

Costello, M. J., Michener, W. K., Gahegan, M., Zhang, Z.-Q., & Bourne, P. E. (2013a). Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology & Evolution*, 28, 454–461.

Costello, M. J., May, R. M., & Stork, N. E. (2013b). Can we name Earth's species before they go extinct? *Science*, 339, 413-416.

Couzin, J. (2006). Truth and consequences. *Science*, 313, 1222–1226.

Craig, I. D., Plume, A. M., McVeigh, M. E., Pringle, J., & Amin, M. (2007). Do open access articles have greater citation impact?: A critical review of the literature. *Journal of Informetrics*, 1, 239–248.

Creative Commons. About the licenses. Available: http://creativecommons.org/licenses/ Accessed 21 February 2014.

Dat: data package management. Available: http://dat-data.com/ Accessed 21 February 2014.

Data Citation Index, Thomson Reuters. Available: http://wokinfo.com/products_tools/multidisciplinary/dci/ Accessed 21 February 2014.

DataCite. Available: http://datacite.org/ Accessed 21 February 2014.

DMPonline. Available: https://dmponline.dcc.ac.uk/ Accessed 21 February 2014.

DMPTool. Available: https://dmp.cdlib.org/ Accessed 21 February 2014.

Drew, B. T., Gazis, R., Cabezas, P., Swithers, K. S., Deng, J., Rodriguez, R., 'et al.' (2013). Lost branches on the tree of life. *PLoS Biol*, 11, e1001636.

Eberhard, W. G. (1985). *Sexual selection and animal genitalia.* Cambridge: Harvard University Press.

Fontaine, B., Perrard, A., & Bouchet, P. (2012). 21 years of shelf life between discovery and description of new species. *Current biology*, 22, 943-944.

Gibney, E. (2013). LHC plans for open data future. *Nature*, 503, 447–447.

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., 'et al.' (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, 11, 156–162.

Hartter, J., Ryan, S. J., MacKenzie, C. A., Parker, J. N., & Strasser, C. A. (2013). Spatially explicit data: Stewardship and ethical challenges in science. *PLoS Biology*, 11, e1001634.

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends*, 57, 280–299.

ImpactStory. Available: http://impactstory.org/ Accessed 21 February 2014.

Lapiedra, O., Sol, D., Carranza, S., & Beaulieu, J. M. (2013). Behavioral changes and the adaptive diversification of pigeons and doves. *Proceedings of the Royal Society B*, 280, 20122893.

Levi, H. W. (1965). Techniques for the study of spider genitalia. *Psyche*, 72, 152-158.

Loon project. Available: http://www.google.com/loon/ Accessed 21 February 2014.

Losos, J. B., Arnold, S. J., Bejerano, G., Brodie, E. D., III, Hibbett, D., Hoekstra, H. E., 'et al.' (2013). Evolutionary biology for the 21st century. *PLoS Biology*, 11, e1001466.

Lücking, R., Mata-Lorenzen, J., & Dauphin, G. L. (2010). Epizoic liverworts, lichens and fungi growing on Costa Rican Shield Mantis (Mantodea: *Choeradodis*). *Studies of Neotropical Fauna and Environment*, 45, 175-186.

Machado, G., Requena, G. S., Buzatto, B. A., Osses, F., & Rosseto, L. M. (2004). Five new cases of paternal care in harvestmen (Arachnida: Opiliones): implications for the evolution of male guarding in the Neotropical family Gonyleptidae. *Sociobiology*, 44, 577–598.

Macaulay Library – The Cornell Lab of Ornithology [http://macaulaylibrary.org]

Moya-Laraño, J., Macías-Ordóñez, R., Blanckenhorn, W. U., & Fernández-Montraveta, C. (2008). Analysing body condition: mass, volume or density? *Journal of Animal Ecology*, 77, 1099–1108.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological reviews of the Cambridge Philosophical Society*, 82, 591–605.

Noorden, V. R. (2014). Elsevier opens its papers to text-mining. *Nature*, 506, 17–17.

Penney, D., McNeil, A., Green, D. I., Bradley, R. S., Jepson, J. E., Withers, P. J., 'et al.' (2012). Ancient ephemeroptera-Collembola symbiosis fossilized in amber predicts contemporary phoretic associations. *PLoS ONE*, 7, 47651.

Peretti, A. V. (2013). Sexual selection in Neotropical species: Rules and exceptions. In R. Macedo, & G. Machado (Eds.), *Sexual selection: perspectives and models from the Neotropics* (pp. 33-52). Amsterdam, ND: Elsevier.

Piwowar, H. A. (2011). Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PLoS ONE*, 6, e18657.

Piwowar, H. A. (2013). Altmetrics: Value all research products. *Nature*, 493, 159–159.

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

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, 1, e175.

Piwowar, H. A., Vision, T. J., & Whitlock, M. C. (2011). Data archiving is a good investment. *Naure*, 473, 285–285.

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): 11-19.

Powell, K. (2012). A lab app for that. *Nature*, 484, 553–555.

Priem, J., Taraborelli, D., Groth, P., Neylon, C. (2010). Altmetrics: A manifesto, (v.1.0). http://alt-metrics.org/manifesto

Ramírez, M. J., Ravelo, A. M., & Lopardo, L. (2013). A simple device to collect, store and study samples of two-dimensional spider webs. *Zootaxa*, 3750, 189.

Raxworthy, C. J., Ingram, C. M., Rabibisoa, N., & Pearson, R. G. (2007). Applications of ecological niche modeling for species delimitation: A review and empirical evaluation using day geckos (*Phelsuma*) from Madagascar. *Systematic Biology*, 56, 907-23.

Roche, D. G., Lanfear, R., Binning, S. A., Haff, T. M., Schwanz, L. E., Cain, K. E., 'et al.' (2014). Troubleshooting public data archiving: Suggestions to increase participation. *PLoS Biol*, 12, e1001779.

Savage, C. J., & Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PLoS ONE*, 4, e7078.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., 'et al.' (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, 6, e21101.

Tummers, B. (2006). DataThief III. http://datathief.org/

Uhl, G, Nessler, S. H., & Schneider, J. M. (2010). Mating plugs and genital mutilation in spiders (Araneae). *Genetica*, 138, 75-104.

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., 'et al.' (2013). The availability of research data declines rapidly with article age. *Current Biology*, 24, 94-97.

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE*, 8, e67332.

Whitlock, M. C. (2011). Data archiving in ecology and evolution: Best practices. *Trends in Ecology & Evolution*, 26, 61–65.

Whitlock, M. C., McPeek, M. A., Rausher, M. D., Rieseberg, L., & Moore, A. J. (2010). Data archiving. *The American Naturalist*, 175, 145–146.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, 61, 726–728.

Winston, J. E. (2007) Archives of a small planet: The significance of museum collections and museum-based research in invertebrate taxonomy. *Zootaxa*, 54: 47-54.

Wolkovich, E. M., Regetz, J., & O'Connor, M. I. (2012). Advances in global change research require open science by individual researchers. *Global Change Biology*, 18, 2102–2110.

Zamir, D. (2013). Where have all the crop phenotypes gone? *PLoS Biol*, 11, e1001595.

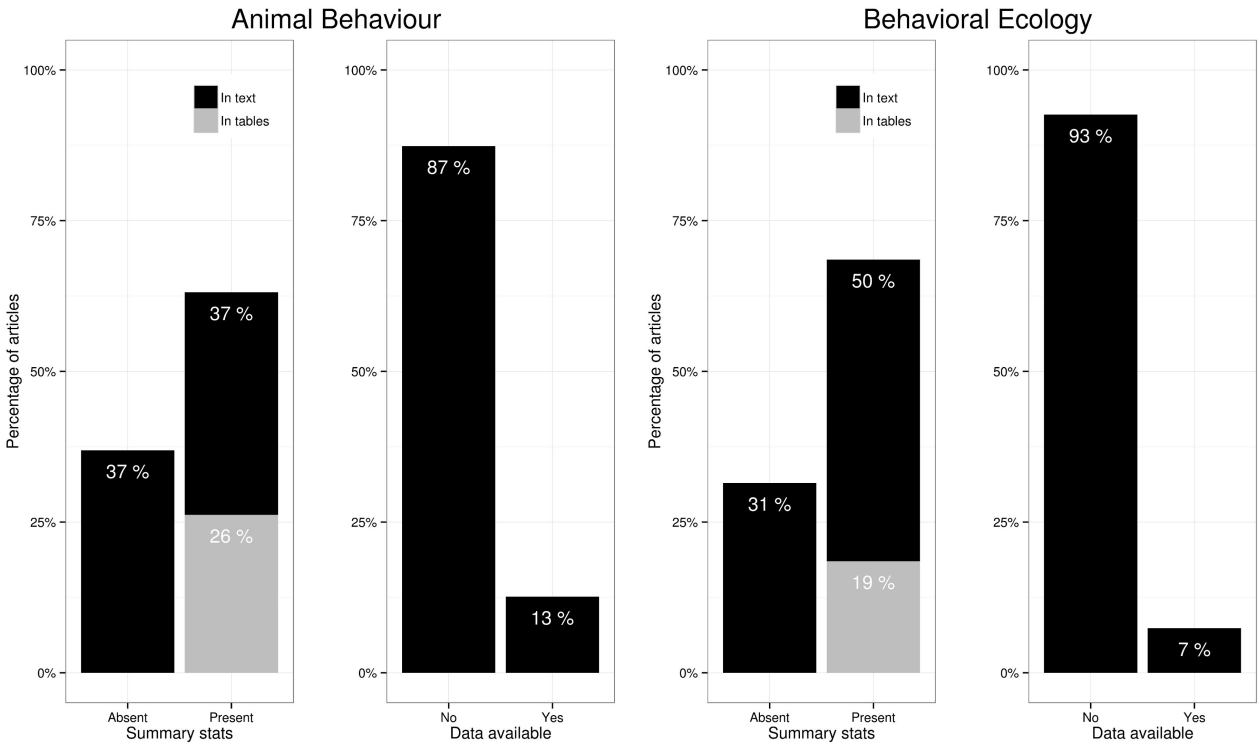

Figure 1. Percentage of randomly chosen articles published in 2013 in Animal Behaviour (103 out of 308) and Behavioral Ecology (54 out of 161) which made their data available. Left bar-plots show whether articles reported summary statistics (e.g., mean/median, standard error and proportions). When summary statistics are present, the black areas show the percentage reported embedded in the text of the article and the gray areas the percentage reported in tables. Right bar-plots show the percentage of articles which raw data were made available in tables or as supplementary material linked to the journal web-site or to databases. See list of sampled articles in Table 1S (DOI: 10.6084/m9.figshare.1003857).

560 **Table 1.** List of repositories suitable for archiving animal behavior data.

Repository	Link	Access to data	Embargo period	Cost	License	File format	File size
Dryad	http://datadryad.org/	Open-access	For selected journals or upon request	Associated fees, with waivers for developing countries	CC0	Any kind	10Gb per data package; Additional fees for bigger packages
fig**share**	http://figshare.com	Open-access; private storage	None	None	CC0 for datasets and CC-BY for media	Any kind	No limit
Macaulay Library	http://macaulaylibrary.org/	Free download for researchers	None	None	Flexible copyright agreement	Video and audio recordings	No limit; consult representative
Zenodo	https://zenodo.org/	Open-access. Have private storage	Yes. Release date set by the authors	None	Author chose among Creative Commons licenses	Any kind	2Gb per data file. Bigger files upon request

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.396v1 | CC-BY 4.0 Open Access | received: 26 May 2014, published: 26 May 2014