Data Management, Archiving, and Sharing for Biologists and the Role of Research Institutions in the Technology-Oriented Age

SÉBASTIEN RENAUT, AMBER E. BUDDEN, DOMINIQUE GRAVEL, TIMOTHÉE POISOT, AND PEDRO PERES-NETO

Data are one of the primary outputs of science. Although certain subdisciplines of biology have pioneered efforts to ensure their long-term preservation and facilitate collaborations, data continue to disappear, owing mostly to technological, regulatory, and ideological hurdles. In this article, we describe the important steps toward proper data management and archiving and provide a critical discussion on the importance of long-term data conservation. We then illustrate the rise in data archiving through the Joint Data Archiving Policy and the Dryad Digital Repository. In particular, we discuss data integration and how the limited availability of large-scale data sets can hinder new discoveries. Finally, we propose solutions to increase the rate of data preservation, for example by generating mechanisms insuring proper data management and archiving, by providing training in data management, and by transforming the traditional role of research institutions and libraries as data generators toward managers and archivers.

Keywords: data management, data archiving, data sharing, data reuse, science policy

Data are one of the primary outputs of science, and extending their lifespan is crucial for several reasons: it helps to foster long-term interdisciplinary research, allows researchers to verify the accuracy of findings, leads to novel insights unforeseen at the time when the data were originally generated, and serves as a support instrument for justifying public funding. Two common examples in ecological and evolutionary research illustrate how data reuse may accelerate the pace of science and technology: (1) publicly available sequences used to build phylogenies and that have become essential to the field of community ecology (Webb et al. 2002) and (2) climatic data combined with species distributions to predict range shifts and explore local to global ecological patterns (Araújo and Rahbek 2006, Kelling et al. 2009, Jetz et al. 2012). Although a growing consensus is emerging about the inherent advantages of sharing data, notable barriers inhibiting the conservation of data remain. These vary from the motivation of individual researchers, appropriate training and technological resources for data management and archiving, and regulatory limitations, as well as ideological barriers affecting the willingness of individuals to share (Tenopir et al. 2011, Whitlock 2011, Hampton et al. 2013, Mills et al. 2015, Whitlock et al. 2016, Wilkinson et al. 2016).

A large number of researchers have probably tried to reuse data without proper documentation and storage (e.g., data recorded in lab notebooks that were not digitized or were recorded on obsolete digital storage media) and realized that in many instances, these data became effectively useless. As such, one major instrument to reduce the alarming rate of the disappearance of data is a proper understanding of what constitutes data management and archiving (Campbell 2009, Vines et al. 2014, Roche et al. 2015, Voytek 2016), as well as the appropriate vehicles toward these goals. *Data management* establishes, at the beginning of a research project, how data will be collected, documented, organized, and preserved (Strasser et al. 2012). *Data archiving* refers to the process of storing no longer actively used data such that they become easily discoverable and reliably retrievable for decades in the future (Whitlock 2011). Finally, *data sharing* is the process of making archived data openly available for reuse (figure 1).

Certain subdisciplines in biological sciences have built strong initiatives toward data archiving (e.g., NCBI GenBank, DataONE, Group on Earth Observations, the National Ecological Observatory Network and the Environmental Data Initiative). These initiatives serve as valuable examples

BioScience 68: 400–411.
doi:10.1093/biosci/biy038 Advance Access publication 24 April 2018

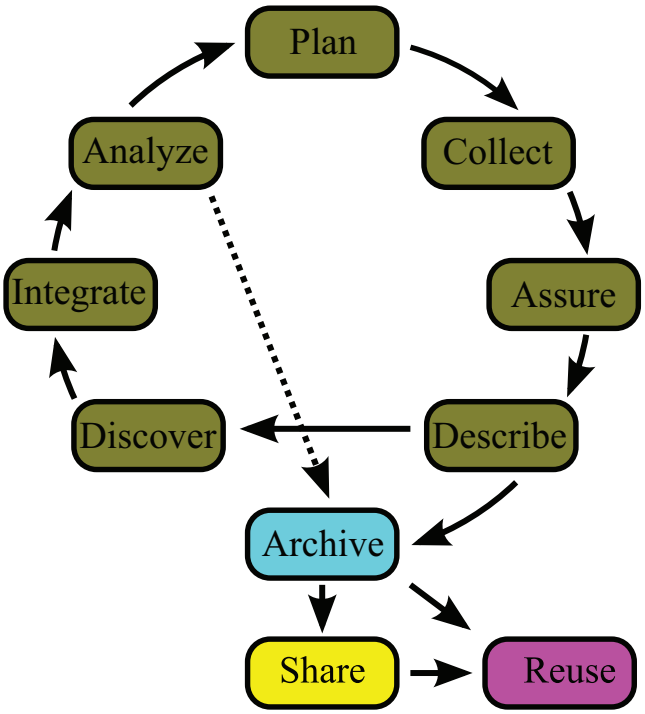

Figure 1. An example of the data life cycle (adapted from* www.dataone.org/data-life-cycle*; Strasser et al. 2012). The data life cycle provides a high-level overview of the stages involved in the successful management and preservation of data for sharing and reuse. It is recognized that some research activities might use only part of the life cycle (e.g., a project focused on primary data collection and analysis might bypass the* discover *and* integrate *steps) or might not follow the linear path depicted here. In this life cycle, data management is the precursor to data preservation (archiving) and should occur once data has been described (raw data with appropriate metadata) and analyzed (summary table). Once this is done, data can be made publicly accessible (shared) or reused.

and resources to promote key concepts, such as data management and training, that we will introduce. In this article, we provide an articulated account for the need of proper data management and archiving, as well as a perspective regarding the main challenges and potential solutions regarding these endeavors. Moreover, based on our experience as members of a large collaborative biodiversity science research center (*www.qcbs.ca*), we believe that building institutional cultures that incentivize and support data management and archiving is essential to preserving data. We illustrate the rise in data archiving and mention the need to better recognize individuals who generate data. In addition, we present specific examples of how data set integration and reuse can lead to new discoveries. Finally, we describe the specific roles that research institutions and libraries can play in order to help researchers manage, archive, and share their data. Although we target mainly ecologists and evolutionary biologists working in academia, many of the issues discussed along with the potential solutions explored in this article are general enough to be applicable to other fields, particularly those that do not deal with sensitive human or medical data.

The importance of conserving data

Digital data have grown immensely in the last 20 years, doubling in size every 2 years and expected to grow to an astonishing 44 zettabytes (4.4×10^{22} bytes) by 2020 (Turner et al. 2014, Voytek 2016); for the sake of comparison, 44 zettabytes is equivalent to the storage capacity of 44 billion laptops. In addition, well over a million scientific articles are published every year (Larsen and Ins 2010), and this number is expected to double every 9 years (Bornmann and Mutz 2015).

Whether publicly funded data should be made public and how individuals that generate data should be recognized when their data are used by others have generated intense debate and potential solutions (Parsons et al. 2010, Mooney and Newton 2012, Piwowar and Vision 2013). In the United States, federal government and agencies have taken note of the problem, and in 2013, the Office of Science and Technology Policy issued a memo requiring US federal agencies to develop a plan to support increased public access to federally funded research results, including data management plans (DMPs, Holdren et al. 2013). More recently, the US National Science Foundation (NSF) began to request that principal investigators list their research *products* rather than *publications*, reflecting the value that data sets, software, and other nontraditional research products contribute to the scientific process. However, despite these initiatives, there is an immediate urgency to act on this critical issue because data continue to disappear.

This point is well illustrated by a recent study that determined that the odds of data being extant decreased by 17% every year after an article was published (Vines et al. 2014). After a certain time period, authors can no longer be easily contacted (e.g., because of broken emails or retirement), data stored on obsolete devices become difficult to retrieve, details (metadata) regarding the study are forgotten, or the data generators are simply reluctant to share. If not properly managed and archived, data will no longer be retrievable even by their own generators (Michener et al. 1997). Clearly, researchers themselves are not the best safekeepers of data for long-term usage and storage (Vines et al. 2013). In addition, Heidorn (2008) reported that most of the data generated in science (about 80%) are never used, reported, or published, an issue referred to as *dark data*. These data may lie dormant in old lab or field notebooks and may never be digitized in a machine-readable format. As such, the volume of data loss is vast and quite evident.

Negative views about data sharing are stalling progress toward data conservation. Opinions vary widely on what constitutes data sharing and how it should be implemented (Whitlock 2011, Mills et al. 2015, Longo and Drazen 2016, Whitlock

et al. 2016). At one end of the sharing spectrum, data may come with no restrictions on reusability, status of the data requester, or research question. At the other end, access to data may be restricted on the basis of what data generators believe is an appropriate reuse of the data and fair recognition to the data generator. In a recent survey among researchers who collect long-term ecological data, Mills and colleagues (2015) found that only 8% of individuals were in favor of uncontrolled data sharing but that the majority were in favor if the principal investigator was involved in the research reusing the data. Many data generating researchers are reluctant to share because of individual costs (Mills et al. 2015, Evans 2016; see table 1 for a summary of benefits and concerns regarding data management, archiving, and sharing). The perception that sharing comes at a high cost to researchers has an insidious consequence: It leads many researchers to adopt a conservative approach regarding data archiving and sharing, perhaps until the debate within the scientific community is settled. As such, it is our view that a focus on this discussion is stalling progress in establishing a culture of proper management and archiving of data.

Individual scientists, governments, journals (via their editorial boards), and institutions are the ones capable of promoting and enforcing data management practices and archiving. However, it is quite clear that the issue of data sharing is rather controversial at the moment. As such, in our views, policies on sharing should occur concurrently or perhaps even be treated as a separate issue. In essence, the concept of *conserving now and sharing later* should be used to convince individual researchers (and institutions) to properly conserve data now. Sharing for future reuse could then be deferred (often referred to as *data embargo*) for when particular data are no longer perceived as strategic by the individuals, groups, or institutions that generated them.

The reproducibility crisis. An apparent rise in the number of scientific retractions (Steen et al. 2013) and the frightening idea that most published findings may simply be wrong (Ioannidis 2005) have led many researchers to believe that science is experiencing a reproducibility crisis (Baker 2016). In many ways, this is the result of scientists failing, consciously or not, to report which parts of the data were used to generate results and which analyses were conducted on the data (selective reporting), thereby biasing interpretation and conclusions (Baker 2016, Parker et al. 2016). Unfortunately, in many cases, we currently do not have the proper means to test, interpret, and reproduce analyses and results reported in published papers (Nekrutenko and Taylor 2012, Open Science Collaboration 2015, Parker et al. 2016). The issue is pressing and needs to be taken seriously if only to ensure that the public continues to trust scientific research (including supporting publicly funded research) and that governmental agencies continue to make science-based decisions to implement policies.

One potential solution to alleviate the reproducibility crisis would be to make data available at the time of submission for reviewers to assess. Although some biological journals have started to request authors to make data available to reviewers, this is usually done when reviewers specifically ask for the data. Assessment of data management and archiving standards during the review process can provide a way to correct errors and make authors check their data and analyses thoroughly before submitting them along with the manuscript. In addition, many journals in ecological and evolutionary research (e.g., *The American Naturalist*, *Molecular Ecology*, *Ecological Monographs*, and the British Ecology Society journals) now request that the data necessary to reproduce the results be archived once the publication is accepted (Whitlock 2011; see also the "Data archiving" section below). Although these represent a virtuous effort to assure data archiving, it is far from solving the issue because the data requested are usually the minimum necessary to reproduce the analyses presented in a paper (Whitlock et al. 2016). Moreover, other solutions will need to be implemented at the same time to address the reproducibility crisis, such as committing to analytical plans prior to collecting data (e.g., *https://cos.io/prereg*).

Defining data

The Open Archival Information System reference model (ISO 2012) defines data as "a reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing" and provides examples of data that include "a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen." Therefore, data can extend beyond digital format, but the preservation of physical artefacts is beyond the scope of the current article. The Office of Science and Technology Policy (OSTP) memo released by the Executive Office of the President in Washington, DC (Holdren et al. 2013), provides guidance on what is not considered data, at least for the purposes of data sharing mandates. These include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer-review reports, and communications with colleagues, as well as physical objects, given that the memo was focused on digital scientific data. Although not intended for sharing, these objects can also be subject to management practices (figure 1; Michener 2015) so that they become appropriately archived for future reference. Consequently, metadata describing each object represent a critical piece of information to ensure that the provenance of data is not lost, in addition to ensuring data interoperability (see the "Giving data a second life through data set integration" section below; Wilkinson et al. 2016, Daraio et al. 2016).

The Data Observation Network for Earth (DataONE 2016, Michener et al. 2011), a global distributed framework of data repositories supporting earth and environmental sciences, can serve as a practical guide to what is being considered as data, at least for the sake of management and archiving. Of the 400,000 publicly available data objects,

Table 1. A summary of the benefits, concerns, and putative solutions regarding data management, archiving, and sharing.

Benefits		Data management	Data archiving	Data sharing
Common understanding of the roles and responsibilities of the researchers involved in the project.		**X**		
Reducing the time associated with data handling and a better version controlling of data.		**X**	x	x
Prevention of the duplication of effort.		**X**	x	x
Insurance to protect against loss.		x	**X**	**X**
Forcing authors to check their data and analyses, thereby preventing errors being committed.		x	**X**	**X**
Less time and effort to retrieve data.		x	**X**	x
Better enforcement of standards regarding quality and formats.		x	**X**	X
Compliance with funder or publisher requirements.		x	x	**X**
Helping solve the reproducibility crisis.		x	x	**X**
Large-scale integration of different sources of information and interoperability with existing databases.		x	x	**X**
Increasing public confidence in science as being open, objective, and transparent.			x	**X**
Investing money more efficiently through data reanalysis rather than spending money collecting new data.			x	**X**
Training tools and data sets for students.				**X**
Leveraging of open databases to make new scientific discoveries.				**X**
Allowing developing countries access to First-World data sets, thereby allowing them to leapfrog technological hurdles.				**X**
Allowing other researchers to work on a problem (i.e., a single individual does not have the time or the creativity to foresee all of the possible ways the data could be analyzed).				**X**
Increasing visibility for the data generator.			x	**X**
Concerns	**Putative Solutions**			
No universal definition of what constitute data.	If the quality and openness of a data set become part of the peer-review process, authors will get feedback and adapt, and each scientific field can come up with their own sets of requirements.	x	**X**	**X**
Lack of proper reward mechanisms.	This should be solved by changing the rules and the culture of evaluations for grants, positions, and promotions. See also "Benefits" above.	**X**	**X**	**X**
It takes time, effort, and money that researchers simply do not have.	It takes less time and effort to manage data as they are being generated than at a later date. Many journals waive data archiving fees.	**X**	**X**	**X**
Sensitive data (e.g., data involving rare species or medical information).	Sensitive information can be anonymized prior to release. Obtain informed consent from the participants.		**X**	**X**
Formats may change, thereby rendering data sets useless.	Use nonproprietary formats.		**X**	**X**
Data can be hacked.	Implement proper security measures, but arguably, this is a real threat with no foolproof solution.		**X**	**X**
Journals and data deposit may disappear along with their data sets.	Archive data on servers that are not tied to a specific journal. Long-term funding of data archives.		**X**	**X**
Researchers have to review data sets, and reviewers may already be overburdened.	Give reviewers proper recognition (see *www.publons.org* as an example of how to start better recognizing reviewers).		**X**	**X**
Without proper knowledge of how the data were collected, the data may be used improperly.	Use a better description of the methods and proper metadata accompanying the data.			**X**
You may end up being scooped with your own data.	You can embargo the data (Mills et al. 2015, Whitlock et al. 2016). Also, this rarely happens, except in very specific disciplines (Mills et al. 2015, Evans 2016, Hendry 2015).			**X**
You may end up being proven wrong with your own data.	No solution. Human error does happen and needs to be recognized and corrected.			**X**
If the new generation of scientists only explores existing data, there will not be interest in training people in acquiring novel data.	Incentivize data collection (i.e., give proper credit to data generators) in order for people to continue collecting data.			**X**
You worked hard for your data, so you should not share it for free (the *research parasite* argument; Longo and Drazen 2016).	No solution, but arguably, data collected with the help of public money should be public, and the nature of research is to build on what has been done before you.			**X**

Note: A large **X** represents a central issue; a small x implies that this is a lesser issue. *Data management* establishes, at the beginning of a research project, how data will be collected, documented, organized, and preserved. It ends once the project is terminated, at which point the data either become obsolete and are deleted or else are archived for posterity. *Data archiving* refers to the process of storing no longer actively used data such that they become easily and reliably retrievable for decades in the future. *Data sharing* is the process of making data openly available to reuse, without restrictions from copyright, patents, or other mechanisms of control.

30% are .csv files, 15% are plain text, 12% are .xml files, 8% are .pdf files, 5% are image files, and 30% are other unknown file types (DataONE 2016). The Dryad Digital Repository (Dryad 2016) archives numerous data sets in ecological and evolutionary research and contains a similar broad scope of data sets.

Ultimately, what are considered data and the resources needed to manage and archive them greatly depends on the discipline (see review by Kratz and Strasser 2014 specifically on this issue). As we mentioned in the previous section, if the quality and level of access of a data set become part of the peer-review process, authors would also receive comments from reviewers about their data, and ideally, each field would evolve toward generating their own set of requirements.

The first step: Planning for data management

Data management should be viewed as a mandatory condition in order for data to be findable, accessible, interoperable, and reusable (FAIR Guiding Principles; Wilkinson et al. 2016), and good data management should take place throughout all stages of the data life cycle (figure 1; Strasser et al. 2012), from planning through collection, assurance, description, preservation, discovery, integration, and analysis (e.g., Michener 2015). For example, the Long Term Ecosystem Research Network (LTER 2017) has pioneered the efficient allocation of limited resources for data management. Notably, its wealth of ecological data across multiples sites in the United States, its long history of data management as an integral core value, and its use of actively maintained and developed metadata standards have made the LTER a successful example of data management supporting scientific research and collaboration. However, challenges such as cross-site and cross-network integration due to the diversity of data types, increased workload due to the implementation of standardized approaches, uneven access to data specialists, and tension between site- and network-specific needs remain.

The first stage of the data life cycle—and one that might be considered an ongoing activity—is planning (figure 1). For example, most US funding agencies now require a data management plan (DMP) as part of the grant submission process to ensure appropriate measures toward the long-term preservation and accessibility of data products arising from federally funded research (see *https://dmptool.org/guidance* for a list of US funders requiring DMPs). A DMP details how to care for the data, including who will be responsible for management and how the data will be documented and archived, especially once the project is completed. Most research sponsors will have specific requirements regarding what should be included in a DMP. For example, the DMPTool (United States; *https://dmptool.org*), DMPonline (United Kingdom, *https://dmponline.dcc.ac.uk*), and DMPAssistant (Canada, *https://assistant.portagenetwork.ca/en*) provide step-by-step tools that enable users to create funder-compliant DMPs. Although variation exists, these templates share a number of primary components and typically include (a) the types of data to be authored, (b) standards that should be applied, (c) roles and responsibilities of individuals involved in the project, (d) access policies, (e) provisions for archiving and preservation, and (f) plans for eventual transition or termination of the data collection involved in a particular study.

Benefits. Regardless of whether these data are ultimately shared publicly, the benefit of creating a data management plan prior to undertaking a research program cannot be overstated (see table 1 on the benefits and concerns regarding data management, archiving, and sharing). Planning for data management can increase efficiency in several ways: by preventing data duplication, reducing the risk of data loss, and reducing the time associated with data handling and error checking (Michener 2015). It also allows researchers to more readily share data with future collaborators (figure 1) and meet journal and other agency requirements. In addition, working on a shared DMP creates a common understanding of the roles and responsibilities of individuals involved in the project: Who will be responsible for and review the DMPs? How will adherence to an established DMP be checked? What process is in place for transferring the responsibility for the data? Who will hold responsibility for the data once the original personnel are no longer involved? Who will hold intellectual property rights, and how will this affect data access? Will the physical location of the data affect its accessibility? Planning in advance will support preservation and access to data in the event of students graduating, collaborators changing institutions or retiring, or the research project simply coming to an end.

Training. Roche and colleagues (2015) suggested that most researchers in ecology and evolution now probably understand their obligation to archive and share data but struggle to do it effectively. Unfortunately, managing and manipulating large volume of heterogeneous information often require specific computational skills that ecologists often do not possess (Poisot et al. 2015). Discipline-specific initiatives that have pioneered collaborative and sharing efforts can serve as potential solutions. For example, DataONE (2016) currently offers training modules on data management that can be incorporated into the current teaching curriculum (*www.dataone.org/education-modules*). Similarly, the National Ecological Observatory Network (NEON) provides data tutorials, workshops, and data management modules (*www.neonscience.org/resources/teaching-modules*), much like the EDI (*https://environmentaldatainitiative.org*) and the CUAHSI (*www.cuahsi.org*). Finally, the success of the Data Carpentry initiative (Teal et al. 2015) also provides strong evidence of the current needs of the community. Generating and reporting appropriate metadata and organizing them in a reusable way across platforms and with appropriate safety controls are all examples of technicalities that the actual curriculum of most biology programs does not cover (see also "The roles of research institutions and

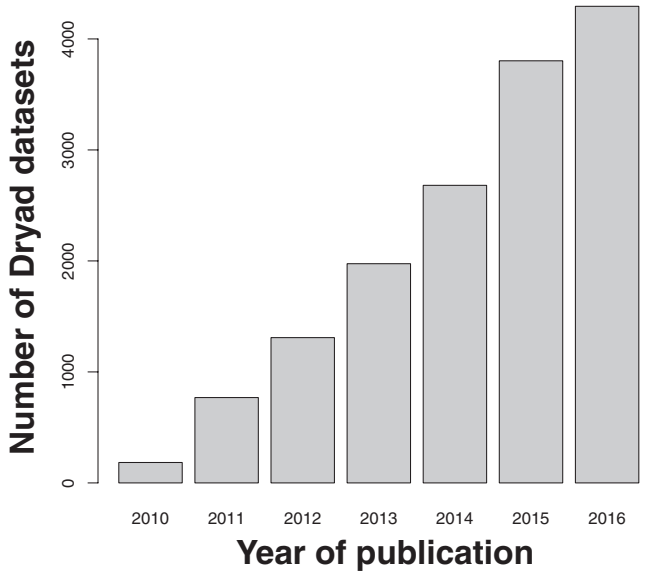

Figure 2. The total number of data packages published in Dryad per year (2010–2016).

their libraries" section). Training is a central component of data management, and research agencies have identified it as a key behavior that a researcher "Learns, stays up to date on and incorporates the basic components of a data, records, and knowledge management process" (NIH 2017).

Data archiving

The number of archiving repositories in science has grown steadily from less than 100 in 2005 to more than 2000 in 2012 (Pinfield et al. 2014). These repositories comprise a small number of large repositories and a large number of small repositories. They are predominantly institutional, multidisciplinary, open access, and English language based (Pinfield et al. 2014). The online tool at *www.re3data.org* helps users identify which storage repository they should target and has identified over 1500 research data repositories, covering a wide range of disciplines and making it the most comprehensive registry available today (Re3Data 2016). With respect to ecological and evolutionary research, repositories have been emerging as *de facto* standards for specific disciplines (e.g., Dryad, GBIF, and NCBI), with DataONE encompassing many discipline-specific initiatives such as the EDI, NEON, or the Europe Long-Term Ecosystem Research Network. In addition, journals now permit researchers to get credited for data generation. Several journals dedicated to data publication exist, and Candela and colleagues (2015) described more than 100 currently existing data journals and journals that publish data papers. Finally, funding invested on open data archiving is arguably very cost effective (Piwowar et al. 2011, Vines et al. 2013). However, maintaining repositories once data have been generated requires continuing support and commitment, and different sources may need to contribute to the financial costs (funding agencies, journal publishers, libraries, or even individual researchers).

The Joint Data Archiving Policy and the role of the Dryad Digital Repository. The Joint Data Archiving Policy (JDAP; see the full description of the policy at *https://datadryad.org/pages/jdap*; Whitlock et al. 2010) describes a requirement that data-supporting scientific publications be openly available and has been adopted by a growing number of journals since 2010. Dryad (2016), a curated resource for data underlying scientific publications, has become the primary repository in ecology and evolution research to comply with the JDAP. Dryad makes data discoverable, freely reusable, and citable. It has been steadily growing from about 100 papers with accompanying data sets in 2010, when a number of journals in ecology and evolution started enforcing the JDAP, to over 4000 in 2016 (figure 2; see Renaut et al. 2018 for data files at *https://datadryad.org*, Renaut 2018 for scripts at *https://zenodo.org*, and SupMat for details). For the 10 biggest journals contributing to Dryad (figure 3), the number of papers with an accompanying Dryad data set has risen steadily (up to approximately 80% in 2015 for *Molecular Ecology*; figure 4). Note that these are likely underestimates of the true rate of data archiving (see SupMat). Noticeably, PLOS ONE has since 2015 become the biggest Dryad contributor (600 data sets in 2015; see figure 3), but this large absolute amount of data package is mainly due to the sheer amount of papers published in PLOS ONE (over 22,000 papers published in 2016). Finally, although JDAP might technically require authors to make data available, without strong enforcement (e.g., through reviewers, editors, and editorial staff making sure that authors comply with the policy), authors often do not comply with the journal policy (Roche et al. 2014, Van Noorden 2014). Indeed, Vines and colleagues (2013) showed that requesting authors to add an explicit data availability statement at the end of their publication was a very efficient way to ensure that authors complied with the journal policy. Finally, badges acknowledging open practices are also effective in order to improve the preservation of data (Kidwell et al. 2016).

Recognizing data generators via data citation. Although the overall rate of data reuse in ecology and evolution appears low (Evans 2016), several examples of novel analyses enabled by this practice are well documented (see "Giving data a second life through data set integration" section). In addition, the rate of data reuse is highly underestimated, in large part because scientists often fail to properly cite their data sources and the current science publication system itself may not be appropriate for data citation. Unfortunately, this is a problem akin to the one faced by software developers, in which most software is improperly recognized and increased citation rates would lead to increased development and sharing (Niemeyer et al. 2016). For example, data sets on Dryad are frequently downloaded, but this only provides circumstantial evidence that they are being reused (Hendry 2015).

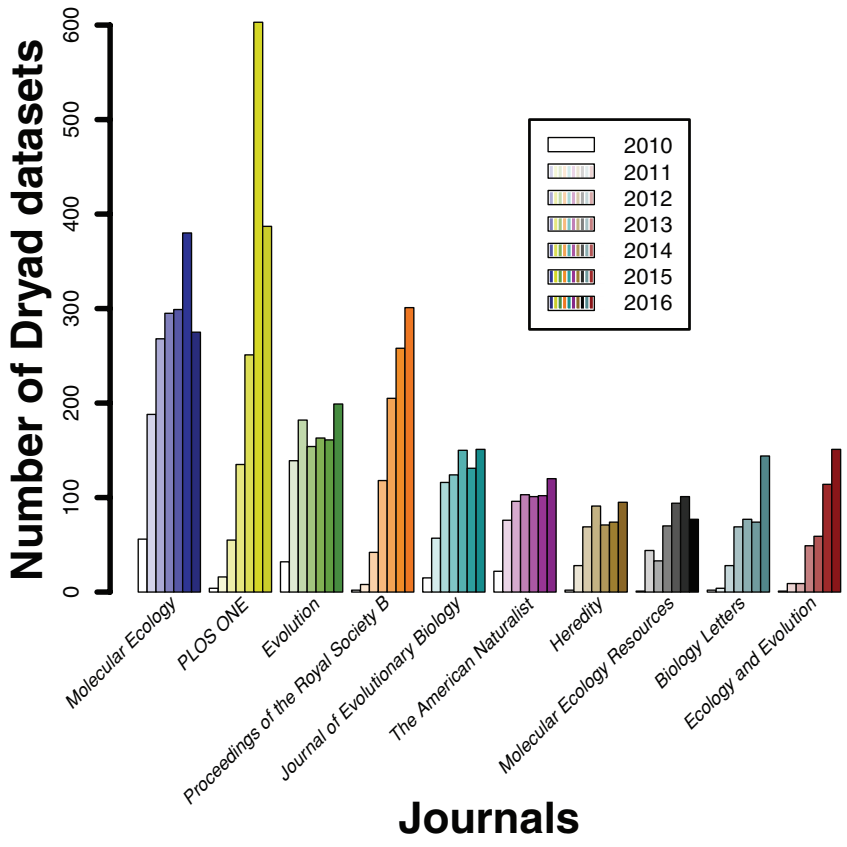

Figure 3. The rise in the number of data packages (2010–2016), broken down for the top 10 biggest journals publishing in Dryad. Most journals show a big rise in 2010–2011, when the Joint Data Archiving Policy (JDAP) was put in place (note that PLOS ONE required authors to provide a data availability statement indicating that all data underlying the findings described in their manuscript are fully available without restriction starting 3 March 2014).

Until publications start explicitly citing data, demonstrating this link will remain difficult.

Data citation will demand effort from at least four fronts. First, data authors must ensure that their archived data are given a unique identifier (such as a digital object identifier, or DOI) and are placed in a publicly accessible and discoverable repository. Second, researchers must consider and be willing to cite data packages and publications as two separate entities. In fact, the idea that data should be published and cited has been advocated for some time (Parsons et al. 2010, Mooney and Newton 2012, Costello et al. 2013), but data citation remains rare in the scientific literature. Third, journals must recognize data packages as a regular scientific publication or product with a DOI. Arguably, citing original data sets, especially from aggregated data, remains problematic. For instance, anyone running a large-scale integrative analysis of biodiversity distribution using GBIF (2016), DataONE (2016), or other large-scale data aggregators faces the problem of reference tractability for these data sets containing thousands of individuals citations. Another issue is proofreading references, which becomes cost prohibitive for thousands of citations. Nevertheless, some journals, such as the Nature Publishing Group's *Scientific Data*, have actually started adding a specific data-citation section (Editorial 2013). Finally, scholarly search engines have also started tracking and indexing data packages, thereby providing data generators with ways of tracking the value (citations) of their work. For example, DataMed (*https://datamed.org*), an NIH initiative, now tracks data packages, in addition to Web of Science's Data Citation Index (*www.wokinfo.com/products_tools/multidisciplinary/dci/*, Thomson Reuters™) and Elsevier's DataSearch (*https://datasearch.elsevier.com/#*), but noticeably not Google Scholar.

Data sharing and reuse lead to new discoveries

Numerous opinions and perspectives have been written specifically on the idea of open science and public data sharing (Tenopir et al. 2011, Goodman et al. 2014, Kratz and Strasser 2014, Roche et al. 2014, Mills et al. 2015). In table 1, we summarize many of the benefits and concerns discussed in the literature. Ultimately, making data available, especially through the use of existing community platforms (such as the ones mentioned in the "Planning for data management" section and in the paragraph below) allows reaching a wider audience, leveraging existing cyberinfrastructure, getting recognition and credit, and enhancing collaborations. Below, we discuss how data reuse can lead to novel discoveries.

Giving data a second life through data set integration. Horizontal data storage (i.e., the aggregation of data sets, keeping their original format) is the first and easiest way to integrate data sets (Jones et al. 2006). However, it will not necessarily be the most useful for advancing research. Alternatively, vertical data storage (i.e., the integration of data sets by constraining information to specific standards) has the potential to open new frontiers of research given how it provides novel information that may not be accessible to a single researcher within a lifetime. The Global Biodiversity Information Facility (GBIF 2016) and NCBI GenBank (Benson et al. 1993) well exemplify the potential that imposing standards can have on generating new results and making data more easily discoverable. Multidimensional storage and the interoperability of databases offer the most potential but require that databases communicate among them and allow different layers of information to be retrieved for a

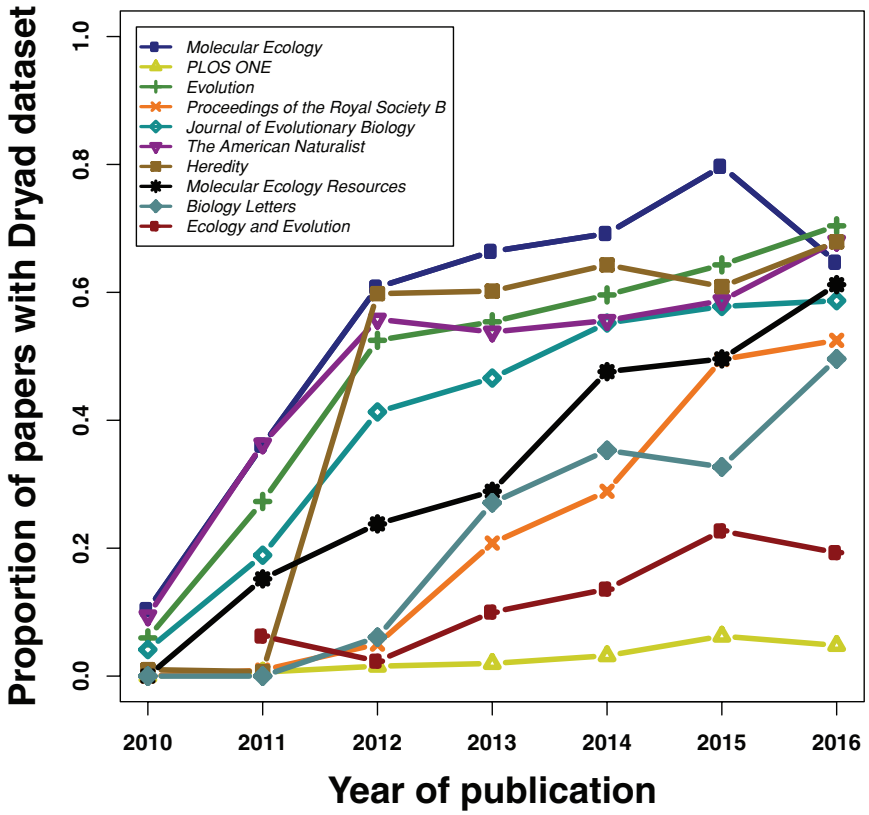

Figure 4. The rise in the proportion of papers, with an accompanying Dryad data set for the top 10 biggest contributing journals (2010–2016).

single data point (Daraio et al. 2016). Although interoperability remains challenging and limits data utility, discipline-specific initiatives exist (e.g., the Global Earth Observation System of Systems, *www.earthobservations.org/geoss.php*; Earthcube, *www.earthcube.org/info/about*; and the National Ecological Observatory Network, *www.neonscience.org*).

Some of the greatest scientific challenges and interests in ecological research are at large spatial scales, integrating a wide range of organisms, systems, or functions into what has been described as macrosystem ecology (Soranno and Schimel 2014). For example, the vertical integration of data sets has recently improved estimates of biodiversity for various groups of organisms, such as trees (Slik et al. 2015), birds (Jetz and Fine 2012), arthropods (Basset et al. 2012), and fishes (Stuart-Smith et al. 2015). This permits a better assessment of the distribution of biodiversity at the global scale as well as improved estimates of extinction rates (Ceballos et al. 2015), including for the rarest and least-documented groups (Régnier et al. 2015). With a comprehensive integration of data sets, we could even envision linking the distribution of biodiversity to information on traits and phylogeny for better assessments of functional and evolutionary distribution mismatch (Devictor et al. 2010) and of extinctions following climate warming (Thuiller et al. 2011). The structure of ecological networks could also be reconstructed across large regions with the combination of trait, species distribution, and biotic interaction data (Albouy et al. 2014).

Although imposing standards helps to make data manageable, discoverable, and reusable, data integration will always remain challenging, simply because, paradoxically, science tends to advance by pushing the boundary of current knowledge standards. As such, the solution likely lies in developing protocols that allow data integration and interoperability rather than settling for a limited number of formats.

The limited availability of suitable large-scale data sets. Analyses of large-scale systems are inherently limited by the availability of suitable data sets: Most data collection results in small-scale, local data, and it is not always clear how these can be used to generate knowledge at larger scales. Collecting exhaustive data sets at large scales can be a daunting effort. However, we envision that ecologists could, in parallel, build on existing databases and aggregate them in a way that allows testing concepts stemming from theory. For example, Poisot and colleagues (2015) illustrated this approach by integrating existing freshwater ecosystem food-web and occurrence data into a synthetic data set, thereby allowing researchers to forecast the structure of stream food webs at the global scale.

Many fields in the biological sciences, genomics being one of the most prominent, rely heavily on data sharing and archiving (e.g., GenBank; Benson et al. 1993). In contrast, some scientific communities have adopted more restrictive practices. For example, the TRY (2016) database, a global archive of curated plant traits, is closed source, and access requires approval of the study by data contributors. As such, it hinders novel analyses and renders the large-scale integration of these data challenging. Even more extreme are data that are not publicly advertised, and in this situation, access to data depends on how connected one is, which is arguably entirely independent of the scientific merit of particular research programs.

The roles of research institutions and their libraries

Although many journals in ecology and evolution as well as funding agencies now mandate that data be archived and made public after a certain period or embargo (Whitlock 2011), a number of issues raised in this article (e.g., what should be archived, enforcement, and dark data) will remain. Obviously, the main obstacle is a lack of policy and guidelines at government levels, but widespread institutional leadership on data management can serve as the path of least

resistance and the next-best option until public policy is put in place. Research institutions, where funding is allocated, managed, and data generated, can establish the necessary framework that includes policies, guidelines, training, and support for data management and archiving. Research institutions already have specific policies and guidelines on how research should be operated, and proper data management could become an accepted norm, like animal use, ethical protocols, and safety codes, which are widespread institutional responsibilities. As we previously advocated, research institutions should play a central role in promoting a culture that supports data management and data archiving (Nature 2009, Royal 2012), and the demonstration that properly managed and archived data have the potential of furthering research is a compelling argument to ensure continuous institutional library funding.

Institutional training in data management. Researchers and their institutions will directly benefit from students leaving properly structured and archived data once they graduate. Institutional practices that allow students to understand their roles in research as well as the benefits of proper data management and archiving are key to data conservation. Students generate most of the data within research institutions, and although training in data management is still in its infancy, this will certainly prove as an additional valuable market skill given the current technology-oriented era. Institutions already provide training in different key aspects of research (e.g., protocols on safety and security, ethics, animal use in research, and the proper management of funds; NRC 2010). Training on data management practices can be implemented via different strategies that best adapt to institutional needs: (a) intensive training sessions and workshops, such as the ones given on animal use in research to both principal investigators and students—these courses could also include ethical components relevant to making data public and deal with the different views regarding sharing; (b) dedicated undergraduate (upper-level) and graduate courses, which can be particularly useful to students working in fields that heavily depend on data integration from public sources (e.g., macroecology, climate change, and genomics); and (c) inclusion in regular course material, particularly those classes dealing with data analyses (e.g., biostatistics and biometrics). A largely unappreciated incentive toward data management is that well-organized data will often facilitate data analyses. The challenge here, however, is to provide training in data management, archiving, and searching (discoverability) that is discipline specific. However, institutions today provide animal-use training by finding internal or external expertise across different taxonomic groups, and similar training strategies could be established to cover differences in data management strategies that vary among subdisciplines.

A model in which data management protocols are detailed in student research proposals would seem quite relevant to promote proper data management and long-term preservation. More specifically, institutions should request that a statement describing how and where the data have been archived accompany dissertations and theses. Given that an explicit data statement at the end of a scientific publication is a highly efficient way to ensure that data are archived (Vines et al. 2013), the same policy should be successful here. Leaving the task to preserve data solely to journals is simply not enough for the many reasons previously stated, if only because in many cases, students will leave academia before publishing their results, in which case, data will rapidly disappear (Vines et al. 2014).

The roles of libraries in data management and archiving. Library science refers to the process of organizing, preserving, and disseminating information. Institutional libraries are well poised to support researchers on data management practices, particularly because data are a chief form of information. The Libraries and Archives Canada, a federal funded institution, states that information management services are "activities undertaken to achieve efficient and effective information management to support program and service delivery; foster informed decision making; facilitate accountability, transparency, and collaboration; and preserve and ensure access to information and records for the benefit of present and future generations." Most researchers (and librarians) would think that this represents a proper description of the roles that libraries can play in their institutions regarding data management and archiving. Training a new generation of librarians to fulfill the role of data managers as well as in assisting researchers and students in this endeavor would certainly fulfill their professional aspirations. Given the current reduction in library services due to the generalized displacement of documents and information to digital formats, librarians would certainly welcome this revitalized purpose. Finally, most libraries already have the long-term vision and infrastructure needed to manage, archive, and generate the protocols for sharing data produced as primary outputs of scientific research (Heidorn 2011).

A movement toward expanding or repurposing institutional libraries so that they serve as institutional vehicles for data management has begun. For example, in 2015, the Canadian Association of Research Libraries launched the Portage Network, dedicated to sharing stewardship of research data and coordinating expertise, services, and technology in research data management (*https://portagenetwork.ca*). The network offers online training resources and assists participants in developing educational materials for their own institutions, and this could meet the challenges of training personnel who have different data management needs, particularly across subdisciplines of biology. Research institutions have also increased the capabilities of libraries to serve their communities in data management. For example, the Data Repository for the University of Minnesota (DRUM; *www.lib.umn.edu/datamanagement/drum*) centralizes data management, archiving, and dissemination. Moreover, institutional libraries are well placed

Box 1. Ten take home messages.

1. The discussion on sharing data is stalling progress on finding solutions to ensure their long-term preservation through management and archiving.
2. Metadata (or data dictionaries) are essential to ensure the long-term safekeeping, reproducibility, interoperability, and reusability of data.
3. Reviewing data should become part of the scientific peer-review process.
4. Data management plans should become a standard practice of scientific projects.
5. Data archiving and sharing have increased steadily since 2010, and many discipline-specific options now exist to archive and share data publicly.
6. Better recognition of data generators (e.g., data citation or coauthorship) is required.
7. Data set integration through sharing leads to reuse and new scientific discoveries.
8. Better training for students and researchers on how to manage data is required.
9. Research institutes granting degrees should enforce explicit data management, archiving, and sharing policies.
10. Libraries should play a central role that includes training in promoting a culture that supports data management and data archiving.

to assist in archiving data that are not digitalized yet (Smith and Rowley 2012).

There are at least three main reasons why libraries need to anchor their efforts with previously established discipline-specific data management, archiving standards, and even existing storage infrastructure. First, it will facilitate establishing the infrastructure (training and resources) and integrating the protocols with existing data archiving systems. Second, the use of existing standards will reduce training and archiving costs, particularly because data storage is free across multiple data archiving systems. Third, using existing standards and archiving systems allows making data more easily discoverable. Finally, librarians can play multiple central roles in data management: (a) guide researchers in choosing appropriate existing public data depositories fitting their views and needs; (b) assist researchers in making their data compliant with existing standards and archiving systems to improve integration and discoverability; (c) assist researchers in searching for data publically available; (d) receive continuous training in best management and archiving practices and available resources—as such, librarians will bring to their institutions the most current approaches regarding this rapidly evolving field; and (e) guide researchers in their strategies for making their data public. For instance, many repositories allow placing embargos on archived data if data generators see preventing sharing as strategic. However, time limits on embargos can be placed, and librarians can manage these even when data are archived in external systems (outside of the institution) so that data eventually become public, assuring long-term archiving and reuse.

In summary, institutions may not necessarily need to invest heavily in infrastructure given the initiatives already available in many subdisciplines of biology, and as Marx (2013) pointed out, cloud storage in public systems is likely the best solution for archiving and sharing.

Conclusions

In this article, our goals were to provide a summary of the issues underlying data management, archiving, and sharing (see also box 1 on our 10 take home messages). Admittedly, we did not cover all aspects and issues pertaining specifically to data generated outside of the academic sector (e.g., governmental research and research performed by the private sector) or involving sensitive human data. Although many of the challenges will be similar, certain legal and regulatory aspects may be specific to researchers working in these sectors.

The discussion about data management, sharing, and open science has progressed immensely in the last decade. Today, most scientists have a minimum level of awareness of the issues presented here, and the publicly available infrastructure necessary for proper data management is rapidly maturing. Specific subdisciplines of biological sciences have provided valuable initiatives and resources to promote collaborative science and which the larger community should embrace. With the experience gained as members of a large biodiversity science center (*www.qcbs.ca*), we believe that currently, the most critical task at hand is to insure that data remain available for future generations and stakeholders. These stakeholders include all groups that contribute to the infrastructure supporting the research activities, in addition to the scientific research community and taxpayers. Although we strongly support data sharing, we recognize that opinions diverge regarding how it should be implemented and optimized. If anything, we hope that the most important message remains undiluted: Managing and archiving data are vital components of research and should

become institutional mandates. The alarming rates at which data are lost need to be reduced, and unlike many contemporary research problems, it appears that we have the means to effectively provide solutions to the issue.

Acknowledgments

We thank Tim Vines, Tim Parker, Louis Bernatchez, and the several anonymous reviewers for comments of earlier versions of this work. We thank the Quebec Centre for Biodiversity Science for providing a platform that allows large-scale collaborative biodiversity science, as well as members of the Centre de la Biodiversité Ecology–Evolution–Genetics Journal club for enlightening discussion around these issues.

The data used to produce figures 2–4 are available at Dryad (*https://doi.org/10.5061/dryad.86634*). The scripts used to analyze data and produce figures 2–4 are available at Zenodo (*http://doi.org/10.5281/zenodo.1185181*).

Funding statement

Research funding was provided by the Quebec Centre for Biodiversity Science, which is a *Fonds de recherche du Québec – Nature et technologies (FRQNT)* strategic research group.

Supplemental material

Supplementary data are available at *BIOSCI* online.

References cited

Albouy C, et al. 2014. From projected species distribution to food-web structure under climate change. Global Change Biology 20: 730–741.

Araújo MB, Rahbek C. 2006. How does climate change affect biodiversity? Science 313: 1396–1397.

Baker M. 2016. Is there a reproducibility crisis? Nature 533: 452–454.

Basset Y, et al. 2012. Arthropod diversity in a tropical forest. Science 338: 1481–1484.

Benson D, Lipman DJ, Ostell J. 1993. GenBank. Nucleic Acids Research 21: 2963–2965.

Bornmann L, Mutz R. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology 66: 2215–2222.

Campbell P. 2009. Data's shameful neglect. Nature 461: 145.

Candela L, Castelli D, Manghi P, Tani A. 2015. Data journals: A survey. Journal of the Association for Information Science and Technology 66: 1747–1762.

Ceballos G, et al. 2015. Accelerated modern human-induced species losses: Entering the sixth mass extinction. Science Advances 1 (art. e1400253).

Costello MJ, Michener WK, Gahegan M, Zhang Z-Q, Bourne PE. 2013. Biodiversity data should be published, cited, and peer reviewed. Trends in Ecology and Evolution 28: 454–461.

Daraio C, et al. 2016. The advantages of an ontology-based data management approach: Openness, interoperability and data quality. Scientometrics 108: 441–455.

[DataONE] Data Observation Network for Earth. 2016. (1 February 2017; *https://search.dataone.org/#profile*)

Devictor V, et al. 2010. Spatial mismatch and congruence between taxonomic, phylogenetic and functional diversity: The need for integrative conservation strategies in a changing world. Ecology Letters 13: 1030–1040.

Dryad. 2016. Dryad Digital Repository. (1 March 2017; *www.datadryad.org*)

Editorial. 2013. Announcement: Launch of an online data journal. Nature 502: 142.

Evans SR. 2016. Gauging the purported costs of public data archiving for long-term population studies. PLOS Biology 14 (art. e1002432–9).

[GBIF] Global Biodiversity Information System. 2016. (1 February 2017; *www.gbif.org*)

Goodman A, et al. 2014. Ten simple rules for the care and feeding of scientific data. PLOS Computational Biology 10 (art. e1003542).

Hampton SE, et al. 2013. Big data and the future of ecology. Frontiers in Ecology and the Environment 11: 156–162.

Heidorn PB. 2008. Shedding light on the dark data in the long tail of science. Library Trends 57: 280–299.

———. 2011. The emerging role of libraries in data curation and e-science. Journal of Library Administration 51: 662–672.

Hendry A. 2015. Archiving primary data (or not). Eco-Evo Evo-Eco. (1 February 2017; *http://ecoevoevoeco.blogspot.ca/2015/12/archiving-primary-data-or-not.html*)

Holdren JP. 2013. Increasing Access to the Results of Federally Funded Scientific Research. Memorandum for the Office of Science and Technology Policy. Executive Office of the President. (21 March 2018; *https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf*)

Ioannidis JPA. 2005. Why most published research findings are false. PLOS Medicine 2 (art. e124).

[ISO] International Organization for Standardization. 2012. The Open Archival Information System Reference model. ISO. (1 March 2017; *www.iso.org/standard/57284.html*)

Jetz W, Fine PVA. 2012. Global gradients in vertebrate diversity predicted by historical area-productivity dynamics and contemporary environment. PLOS Biology 10 (art. e1001292).

Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO. 2012. The global diversity of birds in space and time. Nature 491: 444–448.

Jones MB, Schildhauer MP, Reichman OJ. 2006. The new bioinformatics: Integrating ecological data from the gene to the biosphere. Annual Review of Ecology, Evolution, and Systematics 37: 519–544.

Juffe-Bignoli D, et al. 2016. Assessing the cost of global biodiversity and conservation knowledge. PLOS ONE 11 (art. e0160640–22).

Kelling S, et al. 2009. Data-intensive science: A new paradigm for biodiversity studies. BioScience 59: 613–620.

Kidwell MC, et al. 2016. Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. PLOS Biology 14 (art. e1002456).

Kratz J, Strasser C. 2014. Data publication consensus and controversies. F1000Research 3 (art. 94).

Larsen PO, Ins von M. 2010. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics 84: 575–603.

Lecarpentier D, Wittenburg P, Elbers W. 2013. EUDAT: A new cross-disciplinary data infrastructure for science. International Journal of Digital Curation 8: 279–287.

Longo DL, Drazen JM. 2016. Data sharing. New England Journal of Medicine 374: 276–277.

Lord P, Macdonald A, Lyon L, Giaretta D. 2004. From data deluge to data curation. Pages 371–375 in Cox S, ed. Proceedings of the UK e-Science All Hands Meeting. National e-Science Centre.

[LTER] Long Term Ecological Research Network. 2017. LTER Strategic and Implementation Plan. LTER, National Science Foundation. (9 January 2018; *https://lternet.edu/node/23*)

Marx V. 2013. The big challenges of big data. Nature 498: 255–260.

Michener WK. 2015. Ten simple rules for creating a good data management plan. PLOS Computational Biology 11 (art. e1004525).

Michener WK, Brunt JW, Helly JJ, Kirchner TB, Stafford SG. 1997. Nongeospatial metadata for the ecological sciences. Ecological Applications 7: 330–342.

Michener W, Vieglais D, Vision T, Kunze J, Cruse P, Janée G. 2011. DataONE: Data Observation Network for Earth—Preserving data and enabling innovation in the biological and environmental sciences. D-Lib Magazine 17. doi:10.1045/january2011-michener

Mills JA, et al. 2015. Archiving primary data: Solutions for long-term studies. Trends in Ecology and Evolution 30: 581–589.

Mooney H, Newton M. 2012. The anatomy of a data citation: Discovery, reuse, and credit. Journal of Librarianship and Scholarly Communication 1: 1–14.

Nekrutenko A, Taylor J. 2012. Next-generation sequencing data interpretation: Enhancing reproducibility and accessibility. Nature Reviews Genetics 13: 667–672.

Niemeyer KE, Smith AM, Katz DS. 2016. The challenge and promise of software citation for credit, identification, discovery, and reuse. Journal of Data and Information Quality 7: 16–5.

[NIH] National Institutes of Health. 2017. Data management. NIH Office of Human Resources. (21 March 2018; *https://hr.nih.gov/competency/data-management*)

[NRC] National Research Council. 2010. Guide for the Care and Use of Laboratory Animals, 8th ed. National Academies Press.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. Science 349 (art. aac4716).

Parker TH, et al. 2016. Transparency in ecology and evolution: Real problems, real solutions. Trends in Ecology and Evolution 31: 711–719.

Parsons MA, Duerr R, Minster JB. 2010. Data citation and peer review. Eos 91: 297–298.

Pinfield S, et al. 2014. Open-access repositories worldwide, 2005–2012: Past growth, current characteristics, and future possibilities. Journal of the Association for Information Science and Technology 65: 2404–2421.

Piwowar HA, Vision TJ. 2013. Data reuse and the open data citation advantage. PeerJ 1 (art. e175).

Piwowar HA, Vision TJ, Whitlock MC. 2011. Data archiving is a good investment. Nature 473: 285.

Poisot T, et al. 2015. Synthetic datasets and community tools for the rapid testing of ecological hypotheses. Ecography 39: 402–408.

[Re3Data] Registry of Research Data Repositories. 2016. (1 February 2017; *www.re3data.org*)

Régnier C, et al. 2015. Mass extinction in poorly known taxa. Proceedings of the National Academy of Sciences 112: 7761–7766.

Renaut S. 2018. Seb951/dryad_data_citation: dryad_citation_rate (version 0.92). Zenodo. (21 March 2018; *http://doi.org/10.5281/zenodo.1185181*)

Renaut S, Budden AE, Gravel D, Poisot T, Peres-Neto P. 2018. Data from: Data management, archiving and sharing for biologists and the role of research institutions in the technology-oriented age. Dryad Digital Repository. (21 March 2018; *https://doi.org/10.5061/dryad.86634*)

Roche DG, et al. 2014. Troubleshooting public data archiving: Suggestions to increase participation. PLOS Biology 12 (art. e1001779).

Roche DG, Kruuk LEB, Lanfear R, Binning SA. 2015. Public data archiving in ecology and evolution: How well are we doing? PLOS Biology 13 (art. e1002295).

[Royal] Royal Society Science Policy Centre. 2012. Science as an Open Enterprise. Royal Society. Summary Report no. 02/12.

Slik JWF, et al. 2015. An estimate of the number of tropical tree species. Proceedings of the National Academy of Sciences 112: 7472–7477.

Smith L, Rowley J. 2012. Digitisation of local heritage: Local studies collections and digitisation in public libraries. Journal of Librarianship and Information Science 44: 272–280.

Soranno PA, Schimel DS. 2014. Macrosystems ecology: Big data, big ecology. Frontiers in Ecology and the Environment 12: 3–3.

Statista. 2016. Global GDP (gross domestic product) at current prices from 2010 to 2020 (in billion US dollars). Statista. (1 October 2016; *www.statista.com/statistics/268750/global-gross-domestic-product-gdp*)

Steen RG, Casadevall A, Fang FC. 2013. Why has the number of scientific retractions increased? PLOS ONE 8 (art. e68397).

Strasser C, Cook R, Michener W, Budden A. 2012. Primer on Data Management: What You Always Wanted to Know but Were Afraid to Ask. DataONE.

Stuart-Smith RD, Edgar GJ, Barrett NS, Kininmonth SJ, Bates AE. 2015. Thermal biases and vulnerability to warming in the world's marine fauna. Nature 528: 88–92.

Teal TK, Cranston KA, Lapp H, White E. 2015. Data carpentry: Workshops to increase data literacy for researchers. International Journal of Digital Curation 10: 135–143.

Tenopir C, et al. 2011. Data sharing by scientists: Practices and perceptions. PLOS ONE 6 (art. e21101).

Thuiller W, et al. 2011. Consequences of climate change on the tree of life in Europe. Nature 470: 531–534.

[TRY] TRY Plant Trait Database. 2016. (1 February 2017; *www.try-db.org/TryWeb/Home.php*)

Turner V, Gantz JF, Reinsel D, Minton S. 2014. The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. International Data Corporation.

Van Noorden R. 2014. Confusion over open-data rules. Nature 515: 478–478.

Vines TH, et al. 2013. Mandated data archiving greatly improves access to research data. FASEB Journal 27: 1304–1308.

Vines TH, et al. 2014. The availability of research data declines rapidly with article age. Current Biology 24: 94–97.

Voytek B. 2016. The virtuous cycle of a data ecosystem. PLOS Computational Biology 12 (art. e1005037).

Webb CO, Ackerly DD, McPeek MA. 2002. Phylogenies and community ecology. Annual Review of Ecology and Systematics 33: 475–505.

Whitlock MC. 2011. Data archiving in ecology and evolution: Best practices. Trends in Ecology and Evolution 26: 61–65.

Whitlock MC, et al. 2016. A balanced data archiving policy for long-term studies. Trends in Ecology and Evolution 31: 84–85.

Whitlock MC, McPeek MA, Rausher MD, Rieseberg L, Moore AJ. 2010. Data archiving. American Naturalist 175: 145–146.

Wilkinson MD, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 160018–9.

Sébastien Renaut is affiliated with the Département de Sciences Biologiques of the Institut de Recherche en Biologie Végétale at the Université de Montréal, in Quebec, Canada. Amber E. Budden is affiliated with DataONE at the University of New Mexico, in Albuquerque, New Mexico. Dominique Gravel is affiliated with the Département de Biologie at the Université de Sherbrooke, in Quebec, Canada. Timothée Poisot is with the Département de Sciences Biologiques at the Université de Montréal, in Quebec, Canada. Pedro Peres-Neto is affiliated with the Department of Biology at Concordia University, in Montréal, Québec, Canada. SR, DG, TP, and PPN are also affiliated with the Quebec Centre for Biodiversity Science, in Montréal, Canada. *SR and PPN wrote the manuscript. SR analyzed the data, supervised the editing, produced the figures, and finalized the manuscript. TP and DG drafted the section on data sharing, and AEB drafted the section on planning for data management. All the authors commented on the manuscript and approved its final version.*