GUIDES TO BETTER SCIENCE

# DATA MANAGEMENT

BRITISH ECOLOGICAL SOCIETY

# Contents



BRITISH ECOLOGICAL SOCIETY




**British Ecological Society**
britishecologicalsociety.org
hello@britishecologicalsociety.org

Part of the ***BES Guides to Better Science.*** In this series:
***Peer Review***
***Getting Published***
***Reproducible Code***
***Promoting Your Research***
All are available electronically at **britishecologicalsociety.org/publications/guides-to**

Guide design: Cylinder. Cover image: Adam Seward

# Preface

Good data management is fundamental to research excellence. It produces high-quality research data that are accessible to others and usable in the future. The value of data is now explicitly acknowledged through citations so researchers can make a difference to their own careers, as well as to their fields of research, by sharing and making data available for reuse.

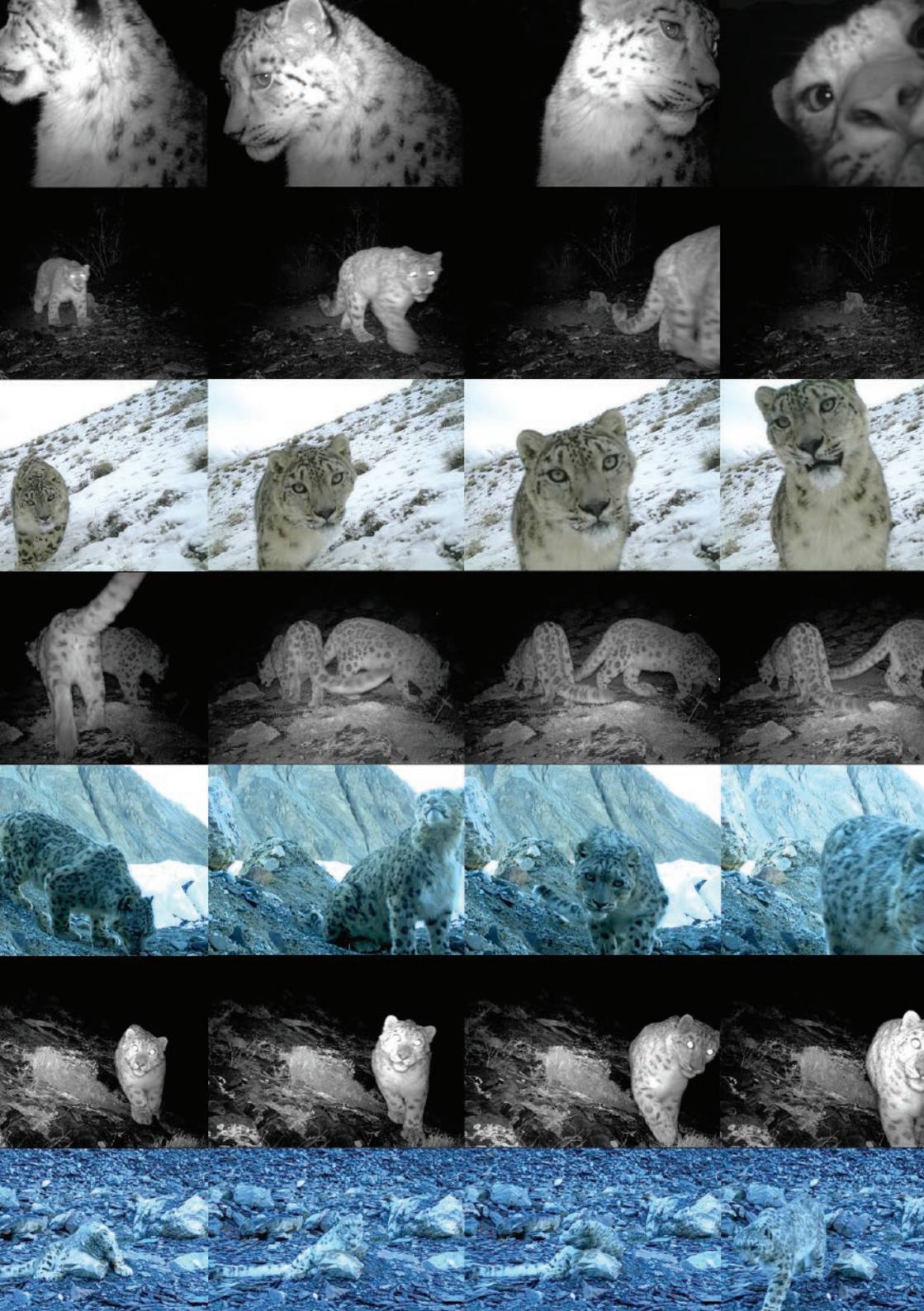
Many of the scientific questions on researchers' agendas now require collaborative research. Data management and sharing is fundamental to this and, consequently, many research funders have come to view data sharing as a priority. This attitude is reflected in their policies and requirements for data management plans during applications for research grants. Digital scientific datasets are a new type of academic output, fundamental for the transparency and cost-efficiency of science today.

Higher education institutions have reacted to this growing trend towards data sharing and management by creating their own policies, guides, tools and training. Publishers have also responded by introducing data archiving policies to ensure that the data behind published research are preserved and made accessible for future use. The British Ecological Society believes that accessibility and preservation of data is important to the fostering of ecological science; in 2011 we formulated a data archiving policy to reflect this and from 2014 introduced a mandate that all data behind published material in our journals be archived.

This guide is designed to help researchers navigate data management firstly by explaining what data and data management are and why data sharing is important, and secondly by providing advice and examples of best practice in data management. The information used in this guide was sourced from various online resources (see **Sources**) and by approaching researchers working with ecology and evolution (see **Acknowledgments**) for real-life examples of the challenges they have faced and lessons they have learned from their own data management experiences.

This new edition (2018) updated in collaboration with GBIF includes sections on international biodiversity data standards and data citation.

***Kate Harrison***
Editor
British Ecological Society



# Introduction

**What are research data?**

Research data are the factual pieces of information used to produce and validate research results. Data can be classified into five categories:

- **Observational**: data which are tied to time and place and are irreplaceable (e.g. field observations, weather station readings, satellite data)
- **Experimental**: data generated in a controlled or partially controlled environment which can be reproduced although it may be expensive to do so (e.g. field plots or greenhouse experiments, chemical analyses)
- **Simulation**: data generated from models (e.g. climate or population modelling)
- **Derived**: data which are not collected directly but generated from (an)other data file(s) (e.g. a population biomass which has been calculated from population density and average body size data)
- **Metadata**: data about data (see **Document**)

A key challenge facing researchers today is the need to work with different data sources. It is not uncommon now for projects to integrate any combination of data types into a single analysis, even drawing on data from disciplines outside ecology and evolution. As research becomes increasingly collaborative and interdisciplinary, this issue will grow in prevalence.

# Introduction


**Fig. 1**. The data lifecycle

**The data lifecycle**

Data often have a longer lifespan than the project they were created for, as illustrated by the data lifecycle (Fig. 1).

Some projects may only focus on certain parts of the lifecycle, such as primary data creation, or reusing others' data. Other projects may go through several revolutions of the cycle. Either way, most researchers will work with data at all stages throughout their career.

Traditionally, researchers were mainly concerned with the early stages of the lifecycle – creating, processing and using. Part of the reason for this was that data sharing and data discovery were much harder to do when data were only on paper and archived in local offices. Now, a combination of technologies that enable data sharing and enabling its discovery and re-use, and the increasing need to combine different datasets to address ecological questions means that preserving and sharing data has become an important part of the scientific process.

# Introduction

It is important to note that research data can and should be used more than once. Once data satisfy the needs of initial collection, open availability and data standardization ensure further uses of data in science and other contexts. Data citation mechanisms enable acknowledgement of the primary contributors of the data. At large scales of analyses, no project or individual, no matter how well-funded, is able to generate all the necessary data anew, and this guide offers advice on data sharing and reuse. For many, data sharing is an uncomfortable first experience, as it is exposes raw, unprocessed evidence, which can compromise academic priority or even career development. The solution comes from deciding on the right moment to share the data: publish your paper first, but remember to share the data with or after the publication of a chapter, book, paper or dissertation.

**Why should I manage data?**

Data management concerns how you plan for all stages of the data lifecycle and implement this plan throughout the research project. Done effectively it will ensure that the data lifecycle is kept in motion. It will also keep the research process efficient and ensure that your data meet all the expectations set by you, funders, research institutions and legislation (e.g. copyright, data protection).

Ask yourself, 'Would a colleague be able to take over my project tomorrow if I disappeared, or make sense of the data without talking to me?' Or even 'Will I be able to find and reuse my own data or recreate this analysis in 10 or 20 years' time?' If you can answer with yes, then you are managing your data well.

Potential benefits of good data management include:

- ensuring data are accurate, complete, authentic and reliable
- increasing research efficiency
- saving time and money in the long run – 'undoing' mistakes is frustrating
- meeting funder requirements
- minimizing the risk of data loss
- preventing duplication by others
- facilitating data sharing
- ensuring data discovery and reuse

**Why should I share my data?**

It is increasingly common for funders and publishers to mandate data sharing wherever possible. In addition, some funding bodies require data management and sharing plans as part of grant applications. Sharing data can be daunting, but

# Introduction

data are valuable resources and their usefulness could extend beyond the original purpose for which they were created. Benefits of sharing data include:

- increasing the impact and visibility of research
- encouraging collaborations and partnerships with other researchers
- maximizing transparency and accountability
- encouraging the improvement and validation of research methods
- reducing costs of duplicating data collection
- advancing science by letting others use data in innovative ways

There are, however, reasons for not sharing data. These include:

- if the datasets contain sensitive information about endangered or threatened species
- if the data contain personal information – sharing them may be a breach of the Data Protection Act in the UK, or equivalent legislation in other countries
- if parts of the data are owned by others – you may not have the rights to share them

During the planning stages of your project you will determine which of your data can and should be shared. Journal data archiving policies recognize these reasons for not sharing. The BES policy, for example, states:

*Exceptions, including longer embargoes or an exemption from the requirement, may be granted at the discretion of the editor, especially for sensitive information such as confidential social data or the location of endangered species.*

Data sharing is one manifestation of a cultural shift towards open science. Other terms that will become more prevalent as this movement grows include:

**Virtual research environments:** a relatively new phenomenon, currently used by only a few universities for collaborative research within the institution. They allow data sharing among colleagues by providing a private virtual workspace for members of a research group to share files, manage workflows, track version control, access resources and communicate with each other.

**Open notebooks:** lab notebooks that are made publicly available online, including all the raw data and any other materials that may be generated in the research project – even 'failed' experiments. They are a transparent approach to research, allowing others to access and feedback on your project in real time, without limitations on

reuse. Open notebooks are not widely used but they are gaining momentum as part of the movement towards open approaches to research practices and publishing.

**Open data:** public data that anyone can use and that are licensed in a way that allows for unrestricted reuse. Advocates of open data are often interested in new computing techniques that unlock the potential of information held in datasets. The term open data came into the mainstream in 2009 when governments, including those of the UK, USA and New Zealand, announced initiatives to open up access to their public information.

**Big data:** a term used to describe extremely large, complex datasets that are difficult to process using traditional methods and require extra computer power. 'Big data' as a concept is more subjective than open data because of this dependence on computers to process them – what seems big today may not seem so big tomorrow when computing technologies are more advanced.

As more and more researchers share their research and work collaboratively, the possibilities of combining open data and big data increase, and the results of this combination have the potential to be very powerful[1]. In fields such as ecology, open and big data could contribute to answering questions on climate change, enable large-scale modeling, and help shape environmental policy.

[1]http://www.theguardian.com/public-leaders-network/2014/apr/15/big-data-open-data-transform-government accessed 10 October 2014.

# Planning data management


Regardless of whether your funder requires you to have a data management or sharing plan as part of a grant application, having such a plan in place before you begin your research project will mean you are prepared for any data management issues that may come your way.

*"I think that having a data management plan is absolutely crucial and planning for archiving and making your data accessible (after a suitable time period if you like) is really important for research in the future. There are now some great resources out there that help you to organize and plan your data collection and make it more usable for yourself as well as others who may wish to use it in the future. Make a plan. Stick to the plan from day one and adapt it as your project evolves and your needs change."*

- Yvonne Buckley, Trinity College Dublin, Ireland

What if things do not go to plan? Good data management is a reflective process and should adapt and respond to changes in circumstance or opportunities that may arise throughout a project. Keep your plan as a 'living document' which is continuously revisited and adapted, if necessary.

Before you start planning:

**Check funder specifications for data management plans.** Each funder will have slightly different requirements but common requirements include: description of the data, quality assurance measures, plans for sharing, restrictions on sharing (if applicable), copyrights and intellectual property rights of data, storage and backup measures, data management roles and responsibilities, and costs of data management.

**Consult with your institution.** Institutions have resources and policies in place that can help throughout the data management process. Policies will state an institution's expectations of good data management as well as outline any issues you need to be aware of when it comes to implementing your plan. Advice and resources for data management on university websites can often be found on the Library or Information Services pages.

# Planning data management

**Consider your budget.** Data management will have its costs and this should be included within the larger budget of your whole research project. You can use the data lifecycle to help price each activity needed throughout data management in terms of people's time or extra resources required. Costing tools may be available from universities and other online data management resources.

**Talk to your supervisor, colleagues and collaborators.** Discuss any challenges they have already faced in their experience – learn from their mistakes.

Key things to consider when planning:

**Time.** Writing a data management plan may well take a significant amount of time. It is not as simple as filling out a template or working through a checklist of things to include. Planning for data management should be thorough before the research project starts to ensure that data management is embedded throughout the research process.

**Design according to your needs.** Data management should be planned and implemented with the purpose of the research project in mind. Knowing how your data will be used in the end will help in planning the best ways to manage data at each stage of the lifecycle (e.g. knowing how the data will be analysed will affect how the data will be collected and organized). Consider using international data standards to ensure that your data will be usable by the most people once the project has finished.

> *"As a student I was always told to plan my analysis well in advance and collect and organize the data toward this. I (and most of my early career colleagues) ignored this advice. Only after spending hours reorganising different datasets did I learn my lesson."*
>
> - Kulbhushansingh Suryawanshi, Nature Conservation Foundation, India

**Roles and responsibilities.** Creating a data management plan may be the responsibility of one single person, but data management implementation may involve various people at different stages. One of the major uses of a data management plan is to enable coordinated working and communication among researchers on a project – with the increase of consortium projects across institutions, this role of planning is a really important way of making sure everyone has a common understanding of what data are being created and used, and under

what terms they are available. During planning it is therefore important to clearly assign roles and responsibilities instead of merely presuming them. Others who may be involved in data management besides you and your collaborators include external people involved in collecting data, university IT staff who provide storage and backup services, external data centres or data archives.

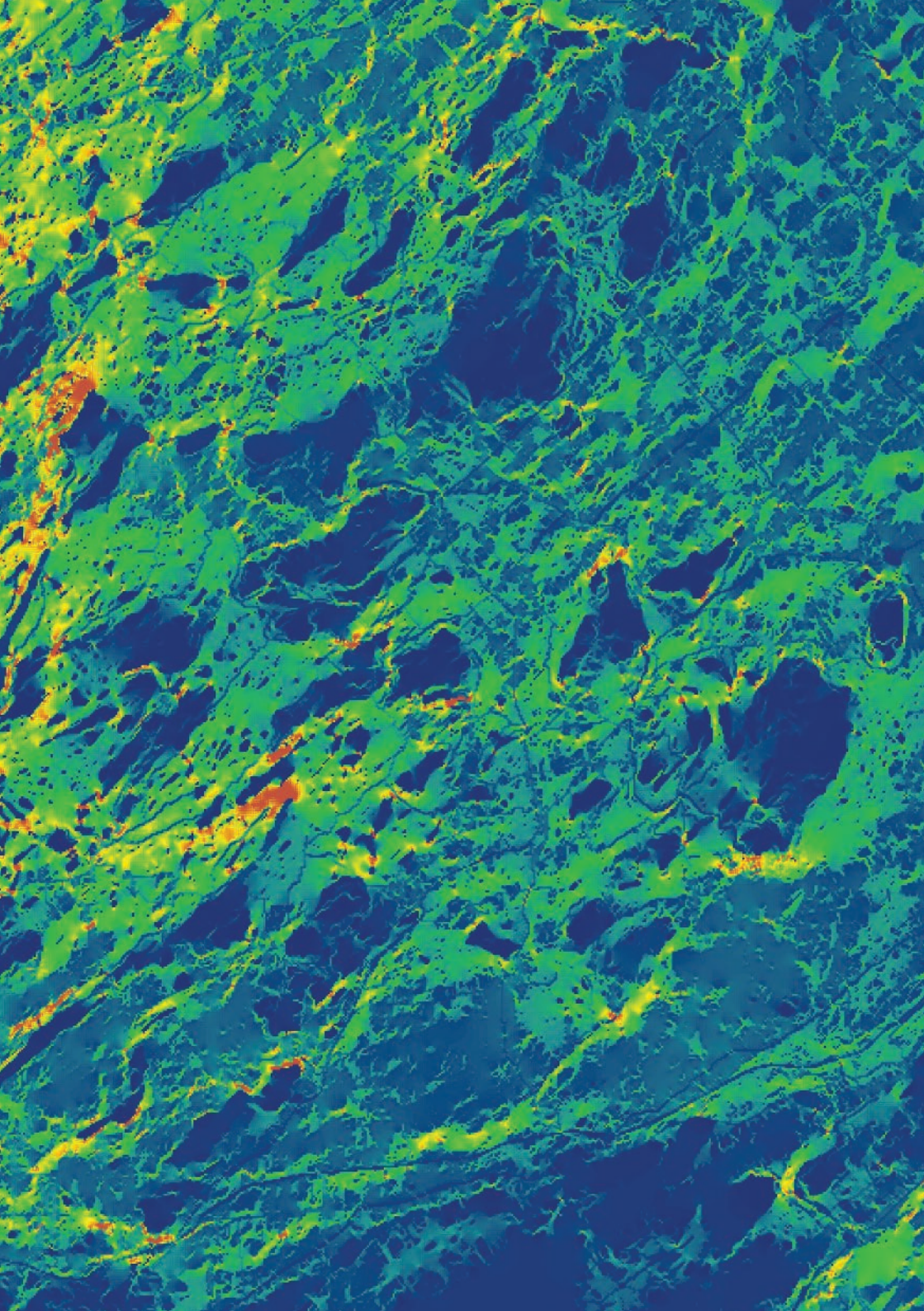
**Review.** Plan how data management will be reviewed throughout the project and adapted if necessary; this will help to integrate data management into the research process and ensure that the best data management practices are being implemented. Reviewing will also help to catch any issues early on, before they turn into bigger problems.

The data management checklist (p14) from the UK Data Archive will help prompt you on the things you need to think about when planning your data management, and enable you to keep on top of your data management once the project has started.

**Data Management Checklist** [2]

- Are you using standardized and consistent procedures to collect, process, check, validate and verify data?
- Are your structured data self-explanatory in terms of variable names, codes and abbreviations?
- Which descriptions and contextual documentation can explain what your data mean, how they were collected and the methods used to create them?
- How will you label and organize data, records and files?
- Will you apply consistency in how data are catalogued, transcribed and organized, e.g. standard templates or input forms?
- Which international data standard is best suited to your data? Will you store your data in that standard or standardize the data exports?
- Which data formats will you use? Do formats and software enable sharing and long-term validity of data, such as non-proprietary software and software based on open standards?
- When converting data across formats, do you check that no data or internal metadata have been lost or changed?
- Are your digital and non-digital data, and any copies, held in a safe and secure location?
- Do you need secure storage for personal or sensitive data?
- If data are collected with mobile devices, how will you transfer and store the data?
- If data are held in various places, how will you keep track of versions?
- Are your files backed up sufficiently and regularly and are backups stored safely?
- Do you know what the master version of your data file is?
- Do your data contain confidential or sensitive information? If so, have you discussed data sharing with the respondents from whom you collected the data?
- Are you gaining (written) consent from respondents to share data beyond your research?
- Do you need to anonymize data, e.g. to remove identifying information or personal data, during research or in preparation for sharing?
- Have you established who owns the copyright of your data?
- Who has access to which data during and after research? Are various access regulations needed?
- Who is responsible for which part of data management?
- Do you need extra resources to manage data, such as people, time or hardware?

[2] UK Data Archive 'Managing and Sharing Data', May 2011, p 35, CC BY-SA 3.0, Copyright 2011 University of Essex



## Creating data


In the data lifecycle creating datasets occurs as a researcher collects data in the field or lab, and digitizes them to end up with a raw dataset.

Quality control during data collection is important because often there is only one opportunity to collect data from a given situation. Researchers should be critical of methods before collection begins – high-quality methods will result in high-quality data. Likewise, when collection is under way, detailed documentation of the collection process should be kept as evidence of quality. Record metadata, the 'data about data' before sampling starts. Document sampling effort, such as area, time, taxonomic targets, any mechanical aids and other important aspects of data collection. This very important envelope of metadata wraps around your data and helps you and others to handle data efficiently. For example, if you need to combine it with other datasets, or when you share the data later.

Key things to consider during data collection:

- logistical issues in the field (see Frida Piper's experience of collecting data on p19)
- calibration of instruments
- taking multiple measurements/observations/samples
- creating a template for use during data collection to ensure that all information is collected consistently, especially if there are multiple collectors
- describing any conditions during data collection that might affect the quality of the data
- creating an accompanying questionnaire for multiple collectors, asking them any questions which may affect the quality of the data
- widening the possible applications of data, and therefore increasing their impact, by adding variables and parameters to data, e.g. wider landscape variables, which will encourage reuse and potentially open up new avenues for research

*"There are very few researchers who have collected 10-year datasets, yet the results that emerge from such data are revealing and impossible to predict from short-term data. Think of associated data that could help you if you had a longer time sequence, then begin to collect those data."*
- Andy Dyer, University of South Carolina Aiken, USA

## Creating data

*"Digitize and organize your data immediately after field collection so it is fresh in your mind and you do not forget about aspects that only the field collector is aware of."* - Roberto Salguero-Gomez, University of Queensland, Australia

*"The project I work on supports a variety of data collection for both long-term data collection, as well as a diverse range of PhD projects. I collect behavioural data through multiple methods, data from laboratory samples and a range of environmental data. The main challenge of such a large-scale research project is due to the fact that there are numerous volunteers and PhD students working together to collect the same data. This means there is a constant need for effective communication and clear ways to record not only the data themselves, but the fact that they have been collected."*
- Cassandra Raby, University of Liverpool, UK

"*If using empirical data collected by someone else, discuss the format of the output and generate a template of the recording spreadsheet for the data prior to recording, including the accompanying explanatory notes and data variable keys. This is particularly important if you are involved in a multi-site experiment – a common recording template provided to all collaborators will make collating the data easier."*
- Caroline Brophy, National University of Ireland Maynooth, Ireland

*"The region where I live and work, Chilean Patagonia, is remote, pristine, isolated and has an often harsh climate. To collect data I usually have to drive on unpaved roads, then hike and climb to the treeline with a cooler box full of ice packs which I use to conserve the tissue samples I collect. As there are no universities in the area, I don't usually have students or assistants to help, and often do this collection alone. Back in the lab there is limited access to basic materials needed to perform chemical analysis, and many of the procedures have never been performed in that lab before, so I must install their protocol for the first time. All of these logistical limitations in the field and the lab mean I must be tough, efficient and independent."*
- Frida Piper, Centro de Investigación en Ecosistemas de la Patagonia, Chile

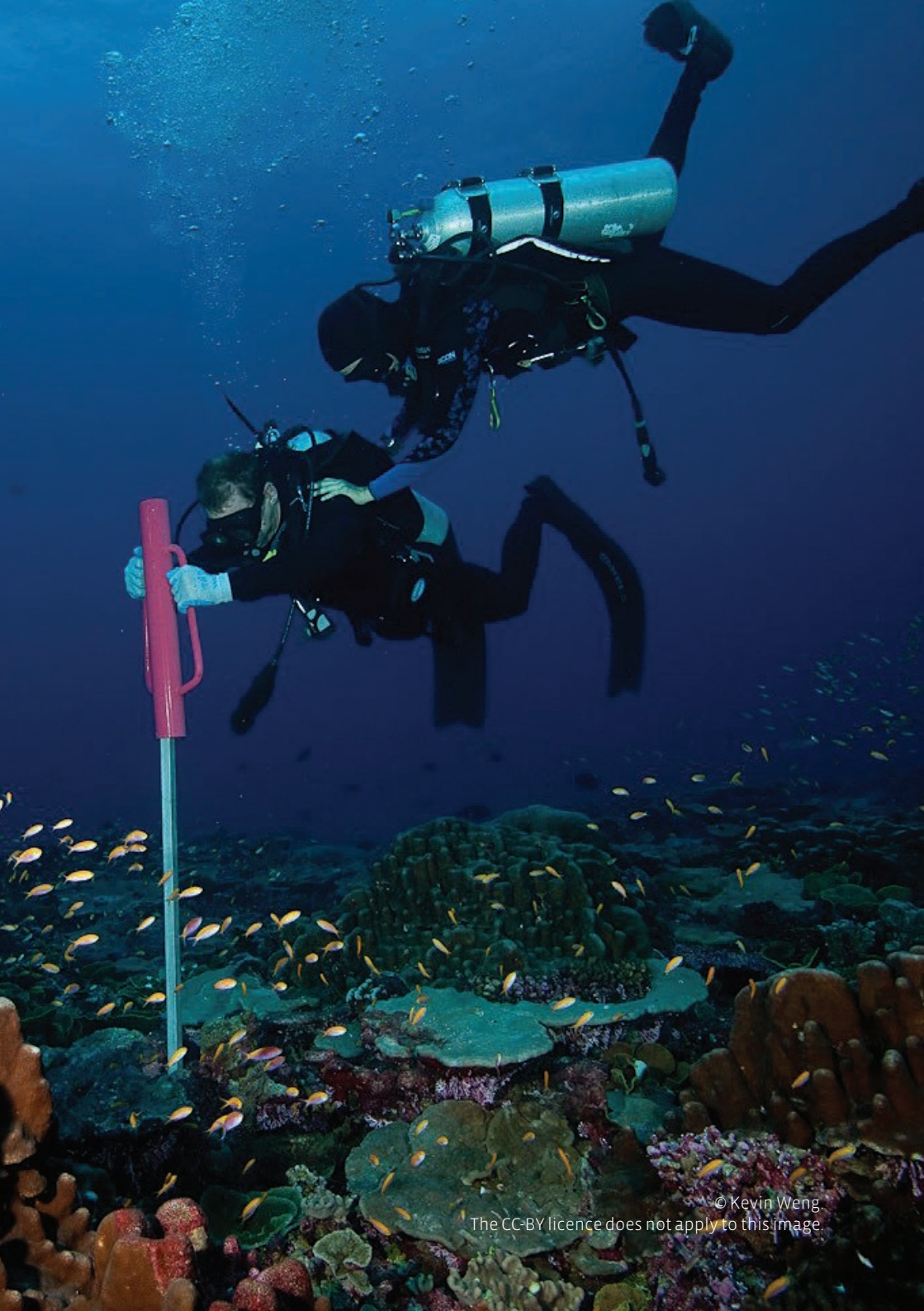
# Creating data

Data may be collected directly in a digital form using devices that feed results straight into a computer or they may be collected as hand-written notes. Either way, there will be some level of processing involved to end up with a digital raw dataset.

Key things to consider during data digitization include:

- designing a database structure to organize data and data files
- using a consistent format for each data file – e.g. one row represents a complete record and the columns represent all the parameters that make up that record (this is known as spreadsheet format)
- atomizing data – make sure that only one piece of data is in each entry
- using plain text characters (e.g. ASCII, Unicode) to ensure data are readable by a maximum number of software programmes
- using code – coding assigns a numerical value to variables and allows for statistical analysis of data. Keep coding simple
- describing the contents of your data files in a 'readme.txt' file, or other metadata standard, including a definition of each parameter, the units used and codes for missing values
- use international data standards – your data are more likely to merge with other data at some point than not. Using international standards helps this process
- keeping raw data raw

*"Despite growing recognition and activity with 'open data', simply posting your data online and making available as-is is not sufficient. The FAIR Data Principles (doi.org/10.1038/sdata.2016.18) offer a framework for ensuring that your data are findable, accessible, interoperable and reusable. The principles also highlight the importance of using accepted biodiversity data standards to ensure that the data you share has real impact for your colleagues. Data in large projects and biodiversity portals come from multiple sources; it is important to consider this when preparing to share and archive your data. In practice, this includes storing your collection dates in the ISO 8601:2004(E) format (YYYY-MM-DD), indicating the coordinate system used for geographic data, and paying attention to metadata. When you are 'done' with your data, take a look at it again from the point of view of an uninformed user: is your data structure self-explanatory? Can the data story be understood from metadata alone without an email from you? If the answer to both questions is yes, you are good to go."*
- Dmitry Schigel, Global Biodiversity Information Facility



# Processing data


Data should be processed into a format that is suited to subsequent analyses and ensures long-term usability. Data are at risk of being lost if the hardware and software originally used to create and process them are rendered obsolete. Therefore, data should be well organized, structured, named and versioned in standard formats that can be interpreted in the future.

Digital information can be altered very easily, so it is important to be transparent in all aspects of processing that have taken place. Careful documentation and using a script will help demonstrate the authenticity of the data.

**File formats.** Data should be written in non-proprietary formats, also known as open standard formats, as far as possible. These files can be used and implemented by anyone, as opposed to proprietary formats which can only be used by those with the correct software installed. The most common format used for spreadsheet data are comma-separated values files (.csv). Other non-proprietary formats include: plain text files (.txt) for text; and GIF, JPEG and PNG for images.

**Filenames and folders.** To keep track of data and know how to find them, digital files and folders should be structured and well organized – this is particularly important when working in collaboration. Use a folder hierarchy that fits the structure of the project (e.g. grouping according to who collected data is only relevant to the collectors, not to those using the data for other things – grouping by data or site may be more relevant) and ensure that it is used consistently. Drawing a folder map which details where specific data are saved may be particularly useful if others will be accessing folders, or if there is simply a lot to navigate.

Filenames should be unique, descriptive, succinct, naturally ordered and consistent. Ideally they should describe the project, file contents, location, date, researcher's initials and version. Computers will add basic properties to a file, such as dates and file type, but relying on these is not best data management practice. Do not use spaces in filenames – these can cause problems with scripting and metadata.

**Quality assurance.** Checking that data have been edited, cleaned, verified and validated to create a reliable masterfile which will become the basis for further analyses. Use a scripting language, such as R, to process your data for quality checking so that all steps are documented.

## Processing data

Assurance checks may include:

- identifying estimated values, missing values or double entries
- performing statistical analyses to check for questionable or impossible values and outliers (which may just be typos from data entry)
- checking the format of the data for consistency across the dataset
- checking the data against similar data to identify potential problems

> *"Communicating the decision process of cleaning data clearly via a scripted audit trial lets others find my mistakes and helps me correct them. Write a script that describes every bit of processing you do. This will mean everything is repeatable and perfectly described; unlike the written list of Excel clicks and copy–pastes you think you carried out, mistakes in scripts can be found and corrected later. You'll never again need a file called "results_final_FINAL_pleaseFINAL.xlsx" because you can keep your analysis and data wrangling in the same script."*
>
> - Will Pearse, University of Minnesota, USA

**Version control.** Once the masterfile has been finalized, keeping track of ensuing versions of this file can be challenging, especially if working with collaborators in different locations. A version control strategy will help locate required versions, and clarify how versions differ from one another.

Version control best practice includes:

- deciding how many and which versions to keep
- using a systematic file naming convention, using filenames that include the version number and status of the file, e.g. v1_draft, v2_internal, v3_final
- record what changes took place to create the version in a separate file, e.g. a version table
- mapping versions if they are stored in different locations
- synchronizing versions across different locations
- ensuring any cross-referenced information between files is also subject to version control

## Processing data

*"I'm a statistician, so I mostly analyse other peoples' data. That can be of all sorts – field observations, genetic data, remote sensing data, almost anything. Getting the data into the format I want can be difficult. Sometimes collaborators can be too helpful, and do all sorts of strange transformations, which I then cannot reproduce. For other projects, I have to pull in data from different sources and getting everything to line up correctly can be a problem.*

*Be clear about the process of going from the raw data to the analyses – use code that can be saved, not a spreadsheet. That way you can check for errors, and make changes more easily. Document what you have done (i.e. write comments in the code) to explain what was done and why. It is important to be able to make the path, from raw data to what is used in the analysis, clear and reproducible."*

- Bob O'Hara, Biodiversität und Klima Forschungszentrum, Germany

# Processing data

**R – a free software environment for statistical computing and graphics**[3]

*"Once a master dataset set has been finalized, all subsequent changes to the data during analysis should be done in R (open source to which all collaborators have access), where changes and analyses can easily be viewed and tracked by all collaborators. In R, multiple datasets of different forms can easily be handled as separate data frames in one single R file and conversions of data form can easily be achieved."*

- Kyle Demes, Simon Fraser University, Canada

*"Working with multiple forms of data is challenging but R can transform your management and analysis of data. My workflows rely on the integration of the R programming language and scripting in the UNIX environment. Scripts and version control of data manipulation and analysis are important and make collaboration transparent and relatively trouble free."*

- Andrew Beckerman, University of Sheffield, UK

*"All modification made to the raw data should be scripted and retained as a record. Use tools like R for your data manipulation because by using R scripts you can keep a record of any changes and manipulations that you do. If you make those changes in Excel you won't remember what you have done 1, 5, 10 years later."*

- Yvonne Buckley, Trinity College Dublin, Ireland

*"R seems to have completely transformed the landscape for data management and analysis. It has limitations as an interpreted language – for instance it is slow for some types of statistical models where a compiled language such as C++ would be faster. However, I have switched over to using R , largely because I can use the R code (with lots of embedded comments) to trace directly from the raw data files (which still need thorough and separate documentation!) to specific results and graphs."*

- Charles Canham, Cary Institute of Ecosystem Studies, USA

[3] http://www.r-project.org/

# Processing data

**Biodiversity Data standards**

Use of common data standards and open-source publishing tools enables data from thousands of different collections and projects to be integrated, discovered and used to support research and policy. These standards are developed by Biodiversity Information Standards (formerly called the Taxonomic Databases Working Group [TDWG]), a not for profit scientific and educational association that is affiliated with the International Union of Biological Sciences. TDWG focuses on the development of standards for exchange of biological/biodiversity data. Darwin Core (DwC) and Darwin Core Archives (DwC-A) are highly recommended standards for publishing data through Global Biodiversity Information Facility (GBIF). Consult with your national GBIF node and GBIF.org for more guidance.

**Why is it important to use TDWG biodiversity standards?**

Data about the world's biodiversity is complex. For centuries, researchers have recorded a wealth of information about the organisms they observe or collect. Public and private institutions around the world manage diverse biodiversity information and with technological advances it is possible to establish connections between all these data. It is important to encourage data holders to publish the richest data possible to ensure their use across a wider range of research approaches and questions. Not every dataset includes information at the same level of detail but sharing what is available is valuable, because even partial information can answer some important questions.

Digital primary biodiversity data can be stored in different formats but for processing they have to be structured. Spreadsheets (or tables) are the easiest and most frequently used form of structuring biodiversity data, where the single row represents a particular occurrence and the columns represent attributes of that occurrence like taxon, locality, observation date or name of the observer. However, as the amount of data increases, the performance of spreadsheet-based programmes deteriorates. A useful alternative is to import data into a database management system. Modern database management systems have a well-developed system of avoiding data redundancy by storing unique data (like taxon names) as separate tables and referencing these data in other tables using numeric identification keys. The system of interconnected tables is known as a relational database. Relational databases significantly increase the amount of data that can be manipulated without reducing performance.


VOLUNTEER
U.S. FISH & WI

# Documenting data


Producing good documentation and metadata ensures that data can be understood and used in the long term. Documentation describes the data, explains any manipulations and provides contextual information – no room should be left for others to misinterpret the data. All documentation requirements should be identified at the planning stages so you are prepared for it during all stages of the data lifecycle, particularly during data creation and processing. This will avoid having to perform a rescue mission in situations where you have forgotten what has happened or when a collaborator has left without leaving any key documentation behind. Metadata tabulates this information for efficient handling by computers.

**Data documentation** includes information at project and data levels and should cover the following:

Project level

- the project aim, objectives and hypotheses
- personnel involved throughout the project, including who to contact with questions
- details of sponsors
- data collection methods, including details of instrumentation and environmental conditions during collection, copies of collection instructions if applicable
- data standards used
- data structure and organisation of files
- data completeness and known gaps
- software used to prepare and read the data
- procedures used for data processing, including quality control and versioning and the dates these were carried out
- known problems that may limit accessibility and data use
- data validation process
- instructions on how to cite the data
- intellectual property rights and other licensing considerations

# Documenting data

Data level

- names, labels and descriptions for variables
- detailed explanation of codes used
- definitions of acronyms or specialist terminology
- reasons for missing values
- derived data created from the raw file, including the code or algorithm used to create them

If a software package such as R is used for processing data, much of the data level documentation will be created and embedded during analysis.

**Metadata** help others discover data through searching and browsing online and enable machine-to-machine interoperability of data, which is necessary for data reuse. Metadata are created by using either a data centre's deposit form, a metadata editor, or a metadata creator tool, which can be searched for online. Metadata follow a standard structure and come in three forms:

- descriptive – fields such as title, author, abstract and keywords
- administrative – rights and permissions and data on formatting
- structural – explanations of e.g. tables within the data

*"Curate your master data file while you create it and document it with metadata at the same time. Creating metadata afterwards is way more painful and you risk misinterpreting your own data, your memory is not usually as good as you think!"*

- Ignasi Bartomeus, Swedish University of Agricultural Sciences, Sweden

# Preserving data


To protect data from loss and to make sure data are securely stored, good data management should include a strategy for backing up and storing data effectively. It is recommended to keep three versions of your data: the original, external/local and external/remote.

**Institutional policies.** These will be in place to regulate methods of data preservation and should be adhered to. Determining a backup and storage procedure will depend on the availability of resources at the institution.

**Backup.** When designing a backup strategy, thought should be given to the possible means by which data loss could occur. These include:

- hardware failure
- software faults
- virus infection or hacking
- power failure
- human error
- hardware theft or misplacement
- hardware damage (e.g. fire, flood)
- backups – good backups being overwritten with backups from corrupt data

The likelihood of each of these will be different, and it may be environment specific (e.g. data being collected in the field may be subject to different risks than data being used across a multi-institutional research team). An ideal backup strategy should provide protection against all the risks, but it can be sensible to consider which are the most likely to occur in any particular context and be aware of these when designing your backup strategy.

Things to consider when drawing up a backup strategy include:

- which files require backup
- who is responsible for backups
- the frequency of backup needed, this will be affected by how regularly files are updated
- whether full or incremental backups are needed – consider running a mix of frequent incremental backups (capturing recent data changes) along with

periodic full backups (capturing a 'snapshot' of the state of all files)

- backup procedures for each location where data are held, e.g. tablets, home-based computers or remote drives
- how to organize and label backup files

**Data storage.** Data storage, whether of the original or backed up data, needs to be robust. This is true whether the data are stored on paper or electronically, but electronic storage raises particular issues. Best practice for electronic storage of data is to do the following:

- use high-quality storage systems (e.g. media, devices)
- use non-proprietary formats for long-term software readability
- migrate data files every two to five years to new storage – storage media such as CDs, DVDs and hard drives can degrade over time or become outdated (e.g. floppy disks)
- check stored data regularly to make sure nothing has been lost
- use different forms of storage for the same data, this also acts as a method of backup, e.g. using remote storage, external hard drives and a network drive
- label and organize stored files logically to make them easy to locate and access
- think about encryption: sensitive data may be regarded as protected while on a password-protected computer, but when backed up onto a portable hard drive they may become accessible to anyone – backups may need to be encrypted or secured too

*"Lots of people solve data storage by buying space on a cloud service such as Dropbox, but they might actually be breaching their contract with their university if they do this, particularly if they store any kind of sensitive material e.g. student-related, patient-related or any kind of social data. People should always consult their institution in terms of implementing not only data storage but all aspects of data management."*

- Rob Freckleton, University of Sheffield, UK

# Preserving data

Data can be stored and backed up on:

- **Network drives** which are managed by IT staff and are regularly backed up. They ensure secure storage and prohibit unauthorized access to files.
- **Personal devices** such as laptops and tablets are convenient for short-term, temporary storage but should not be used for storing master files. They are at high risk of being lost, stolen or damaged.
- **External devices** such as hard drives, USB sticks, CDs and DVDs are often convenient because of their cost and portability. However, they do not guarantee long-term preservation and can also be lost, stolen or damaged. High-quality external devices from reputable manufacturers should be used.
- **Remote or online services** such as Dropbox, Mozy and A-Drive use cloud technology to allow users to synchronize files across different computers. They provide some storage space for free but any extra space or functions will have to be bought.
- **Paper!** If data files are not too big, do not overlook the idea of printing out a paper copy of important ones data files as a complement to electronic storage. It may not be convenient, but ink on paper has proven longevity and system independence (as long as you can remember where you put it)!

*"I recently had a request for the raw data from a 25-year-old experiment and was able to find the floppy disk but then spent a day trying to figure out how to translate from an old software format into something readable. The request was for data that didn't make it into the published paper but that we had in fact collected. It felt good to be able to offer it up for use in someone else's work."*

- Charles Canham, Cary Institute of Ecosystem Studies, USA

# Sharing data


Research data can be shared in many ways and each method of sharing will have advantages and disadvantages. Ways to share data include:

- using a disciplinary data centre such as Dryad or GenBank
- depositing data in your research funder's data centre
- depositing data in university repositories
- sharing standardized data exports though national biodiversity portals and GBIF
- making data available online via open notebooks or project websites
- using virtual research environments such as SharePoint and Sakai

The BES data archiving policy, which mandates that all data used to support the results in papers published in its journals be archived in a suitable repository that provides 'comparable access and guarantee of preservation'[4], encourages authors to pick a repository best suited to their type of data and is most useful to the community most likely to access their data. You may consider following the ICSU-World Data System Data Sharing Principles[5] which outline that data should be shared openly in a timely manner, with the fewest restrictions possible and used with proper citation.

**Data repositories.** Archiving your data in a repository is a reliable method of sharing data. Data submitted to repositories will have to conform to submission guidelines, which may restrict which data you share via the repository. However, the advantages of sharing data through these centres include:

- assurance for others that the data meet quality standards
- guaranteed long-term preservation
- data are secure and access can be controlled
- data are regularly backed up
- chances of others discovering the data are improved
- citation methods are specified
- secondary usage of the data is monitored

[4]http://www.britishecologicalsociety.org/publications/journal-policies/#data accessed 9 October 2018
[5]www.icsu-wds.org/services/data-sharing-principles accessed 9 October 2018

# Sharing data

Longitudinal datasets that span many years are important in ecology and evolution. Journals mandating that authors archive their data only guarantees the preservation of the data used in a particular paper, but researchers should be aware of the value of archiving and sharing large datasets to drive discovery. Sharing large datasets has not been common practice in fields such as ecology. However, as trust in ethical guidelines, journal and funder requirements, and community expectations with regards to accessing and correctly citing others' data grow, progress can be made towards a more open access data future.

**Data papers.** A data paper is a peer-reviewed paper published in a scholarly journal describing a particular dataset or a group of datasets. The primary purpose of a data paper is to present the metadata and describe the data and the circumstances of their collection, rather than to report hypotheses and conclusions. By publishing a data paper, you will receive credit through indexing and citation of the published paper, in the same way as with a research paper.

**Publishing biodiversity data through the Global Biodiversity Information Facility (GBIF).** The Global Biodiversity Information Facility (www.gbif.org) is an open-data research infrastructure, funded by governments, aimed at providing anyone, anywhere with access to data about all types of life on Earth. Coordinated through a Secretariat in Copenhagen, Denmark, GBIF enables data-holding institutions around the world to share information about where and when species have been recorded. This knowledge derives from many sources, including museum specimens dating back decades or centuries, current research data and monitoring programmes, as well as volunteer recording networks and citizen science initiatives.

By encouraging the use of common data standards and open-source publishing tools, GBIF enables data from thousands of different collections and projects to be integrated, discovered and used to support research and policy. Data published through GBIF can be freely accessed at the global level through GBIF.org and associated web services, as well as through national and thematic portals making use of the shared infrastructure.

Through its network of national, regional and thematic nodes, (see www.gbif.org/the-gbif-network), GBIF also acts as a collaborative community, sharing skills and best practices to encourage the widest possible participation.

## Sharing data

By sharing data using GBIF-compatible tools, researchers and institutions can:

- Add value to the data by enabling its re-use across a wide range of research fields
- Fill geographic, taxonomic and temporal data gaps, thus advancing biodiversity knowledge both within the region and beyond
- Provide visibility for natural history collections and research projects, including individuals involved at all levels such as field collection, identification, curation and data management
- Track the uses and applications of data through citation in research and metrics on data downloads
- Meet obligations on data management and access increasingly required by funding agencies and public authorities

## Reusing data


All aspects of data management lead up to data discovery and reuse by others and yourself. Intellectual property rights, licenses and permissions, which concern reuse of data, should be explained in the data documentation and/or metadata. At this stage of the lifecycle it is important to state your expectations for the reuse of your data, e.g. terms of acknowledgement, citation and coauthorship. Likewise, it becomes the responsibility of others to reuse data effectively, credit the collectors of the original data, cite the original data and manage any subsequent research to the same effect.

When requesting to use someone else's data it is important to clearly state the purpose of the request, including the idea you will be addressing and your expectations for coauthorship or acknowledgement. Coauthorship is a complex issue and should be discussed with any collaborators at the outset of a project.

Increasing openness to data and ensuring long-term preservation of data fosters collaboration and transparency, furthering research that aims to answer the big questions in ecology and evolution. By implementing good

data management practices, researchers can ensure that high-quality data are preserved for the research community and will play a role in advancing science for future generations.

Data openness

- increases the efficiency of research
- promotes scholarly rigor and quality of research
- enables tracking of data use and data citation through DOIs
- expands the spectrum of academic products through data papers
- enables researchers to ask new research questions
- enhances collaboration and community-building
- increases the economic and social impact of research
- supports international conventions and requirements from funding agencies

Citing data. Data accessed from data portals is often free and open, but is not free of obligations. Read and respect the data use agreements of the data portals you use in your research, and follow the citation guidelines of the data portal and instructions to authors in your journal. Whenever possible, use a DOI to refer to the unprocessed or downloaded data, and to the processed and archived version. Good citation practices ensure scientific transparency and reproducibility by guiding other researchers to the original sources of information. They also reward data-publishing institutions and individuals by reinforcing the value of sharing open data and demonstrating its impact to their stakeholders and funders. Datasets published through GBIF and other portals are authored electronic data publications and, as such, should be treated as first-class research outputs and correctly cited.

# Sources and Further Reading

All online sources were accessed 9 October 2018.

BES, Data Archiving Policy www.britishecologicalsociety.org/publications/journal-policies

Biodiversity information management and reporting guidelines for South-East Europe 2018. GIZ Open Regional Fund for South-East Europe – Biodiversity (ORF-BD) www.balkangreenenergynews.com/wp-content/uploads/2018/03/BIMR_ENG_publication_Final-Preview.pdf

EPSRC, Clarifications of EPSRC expectations on research data management www.epsrc.ac.uk/files/aboutus/standards/clarificationsofexpectationsresearchdatamanagement

GBIF: The Global Biodiversity Information Facility 2018 Quick guide to publishing data through GBIF.org www.gbif.org/publishing-data

GBIF: The Global Biodiversity Information Facility 2018 Data standards www.gbif.org/standards

DataONE Best Practices Primer www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf

DCC, How to Develop a Data Management and Sharing Plan www.dcc.ac.uk/resources/how-guides/develop-data-plan

MANTRA, Research Management Training mantra.edina.ac.uk

Norman, H. 'Mandating data archiving: experiences from the frontline' Learned Publishing 27 S35-S38

UK Research and Innovation, Common Principles on Data Policy www.ukri.org/funding/information-for-award-holders/data-policy

UK Data Archive, Managing and Sharing Data (2011) www.data-archive.ac.uk/media/2894/managingsharing.pdf

# Acknowledgements

This booklet would not have been possible without contributions from: Peter Alpert, Andrea Baier, Liz Baker, Ignasi Bartomeus, Andrew Beckerman, Caroline Brophy, Yvonne Buckley, Rosalie Burdon, Charles Canham, Tim Coulson, Kyle Copas, Kyle Demes, Stéphane Dray, Andry Dyer, Rob Freckleton, David Gibson, Erika Newton, Bob O'Hara, Catherine Hill, Tim Hirsch, Will Pearse, Nathalie Pettorelli, Frida Piper, Cassandra Raby, Andrew Rodrigues, Roberto Salguero-Gomez, Dmitry Schigel, Kulbhushansingh Suryawanshi, Phil Warren and Ken Wilson.

## Image credits

p2: Norwegian University of Life Sciences /Snow Leopard Foundation Pakistan

p3: Danielle Green

p6: Markku Larjavaara

p8: David Bird

p9: Kara-Anne Ward

p10: Ute Bradter

p13: Benjamin Blonder

p15: © Jeremy Holloway

p16: Image provided by Koen and Walpole using Circuitscape

p18 Tomáš Václavík

p20: © Kevin Weng.
The CC-BY licence does not apply to this image.

p23: Oliver Hyman

p26: Hannah Grist

p30: Adam Seward

# PUBLICATIONS

Proud to partner with Ecology and Evolution Open Access

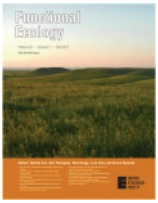
## Journal of Ecology

**journalofecology.org**
**@jecology**

High-impact, broad reaching articles on all aspects of plant ecology (including algae), in both aquatic and terrestrial ecosystems.

## Functional Ecology

**functionalecology.org**
**@funecology**



High-impact papers that enable a mechanistic understanding of ecological pattern and process from the organismic to the ecosystem scale.

## Journal of Applied Ecology

**journalofappliedecology.org**
**@jappliedecol**

Novel, high-impact papers on the interface between ecological science and the management of biological resources.

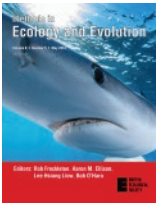
## Journal of Animal Ecology

**journalofanimalecology.org**
**@animalecology**

Publishing the best animal ecology research that develops, tests and advances broad ecological principles.

## Methods in Ecology and Evolution

**methodsinecologyandevolution.org**
**@methodsecolevol**



Promotes the development of new methods and facilitates their dissemination and uptake by the research community.

## People and Nature

**people-and-nature.org**
**@PaN_BES**

A broad-scope open access journal publishing work from across research areas exploring relationships between humans and nature. **Now open for submissions!**

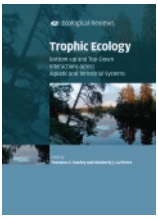
## Guides to Better Science

**bit.ly/GuidesToBetterScience**

Promoting research excellence across a range of topics including peer review, data management, reproducible code and getting published. These free guides contain practical tips for researchers all over the world.

## Ecological Reviews

**bit.ly/EcologicalReviews**



Books at the cutting edge of modern ecology, providing a forum for current topics that are likely to be of long-term importance to the progress of the field.