Data rescue: saving environmental data from extinction

Ellen K. Bledsoe[1,2,†,*], Joseph B. Burant[1,3,4,†,*], Gracielle T. Higino[1,5,†], Dominique G. Roche[1,6], Sandra A. Binning[1,4], Kerri Finlay[1,2], Jason Pither[1,7], Laura S. Pollock[1,3], Jennifer M. Sunday[1,3], Diane S. Srivastava[1,6,*]

Author affiliations

1. The Living Data Project, Canadian Institute of Ecology and Evolution, Vancouver, British Columbia, Canada
2. Department of Biology, University of Regina, Regina, Saskatchewan, Canada
3. Department of Biology, McGill University, Montreal, Quebec, Canada
4. Département de Sciences Biologiques, Université de Montréal, Montréal, Québec, Canada
5. Department of Zoology and Biodiversity Research Centre, University of British Columbia, Vancouver, British Columbia, Canada
6. Department of Biology and Institute for Environment & Interdisciplinary Science, Carleton University, Ottawa, Ontario, Canada
7. Department of Biology and Okanagan Institute for Biodiversity, Resilience, and Ecosystem Services, University of British Columbia, Kelowna, British Columbia, Canada

† These co-authors contributed equally to this work and reserve the right to prioritise their names in the publication list on their CV.

* Corresponding authors: ellen.bledsoe@weecology.org (EKB), joseph.burant@mcgill.ca (JBB), srivast@zoology.ubc.ca (DSS)

Running Headline

Data rescue: saving environmental data

Author contributions

EKB, JBB, DGR, and DSS proposed the initial idea for the manuscript; all authors contributed to developing the methods of data rescue we describe and subsequent discussions about the paper. EKB, JBB, and GTH wrote the first draft. DSS created the first draft of the visual. All authors revised the manuscript for publication. The authors declare no competing interests.

Abstract

Historical and long-term environmental datasets are imperative to understanding how natural systems respond to our changing world. Although immensely valuable, these data are at risk of being lost unless actively curated and archived in data repositories. The practice of data rescue, which we define as identifying, preserving, and sharing valuable data and associated metadata at risk of loss, is an important means of ensuring the long-term viability and accessibility of such datasets. Improvements in policies and best practices around data management will hopefully limit future need for data rescue; these changes, however, do not apply retroactively. While rescuing data is not new, the term lacks formal definition, is often conflated with other terms (i.e., data reuse), and lacks general recommendations. Here, we outline seven key guidelines for effective rescue of historically-collected and unmanaged datasets. We discuss prioritisation of datasets to rescue, forming effective data rescue teams, preparing the data and related metadata, and archiving and sharing the rescued data. In an era of rapid environmental change, the best policy solutions will require evidence from both contemporary and historical sources. It is, therefore, imperative that we identify and preserve valuable, at-risk environmental data before they are lost to science.

Keywords

Data archiving, historical data, long-term ecological data, long-term studies, open data, open science, reproducibility, transparency

Total word count (excluding title page contents and supporting information) = 7,436
- Abstract word count = 199
- Main text word count (excluding abstract, figure captions, boxes, and references) = 4,390

Figures: 2
- Figure 1 caption = 83
- Figure 2 caption = 80

Boxes: 2
- Box 1 = 327 words
- Box 2 = 918 words, 2 images

References = 50 (1,439 words)
Supporting items: 2

Why Rescue Data?

Data are among the most valuable outputs of research and scholarship; beyond helping answer important questions, they inform new lines of inquiry, new testable hypotheses, and future data collection efforts. Observational and experimental data derived from ecology, evolution, conservation and environmental sciences (hereafter, environmental data) are essential to establishing historical trajectories of ecosystems ("baselines") [1], understanding how species and communities respond to environmental change [2], and designing and evaluating the outcomes of management efforts [3]. While data collection is often targeted to particular populations, communities, or locations, the reuse (i.e., aggregation, collation, and synthesis) of data from different contexts is essential to establishing broader ecological knowledge and informing conservation management [4]. Yet, despite their high value, data are often misplaced, filed away, or otherwise rendered unusable, often through poor data management practices [5]. In their unusable and "at-risk" state, these data represent an egregious waste of resources expended on their collection (Box 1) [6]. Languishing data, however, also offer an enormous opportunity. **Data rescue**—defined here as the identification, preservation, and sharing of valuable data and associated metadata at risk of loss—has the potential to realise substantial benefits for society, especially considering the crucial roles that baseline data play in informing management and policy decisions. The ultimate goal of data rescue is to make previously inaccessible or poorly preserved data available for (re)use, ideally through archiving them in a permanent, publicly accessible, and reusable format.

Data rescue is particularly important in the environmental sciences for three reasons. First, because environmental processes are context-dependent, they often have historical components. Such records are essential in understanding the trajectory of environmental change and guiding policy to mitigate or adapt to this change [7]. For example, information obtained by rescuing salmon samples collected in the early 20th century dramatically changed our understanding of how salmon stocks have declined over the last century [8]. Second, environmental datasets are often small and local, constrained by both organismal-level data collection and the fine spatial scale of many of the underlying processes. Therefore, to obtain powerful tests of theory and the generality of mechanisms across heterogeneity in ecosystems and species, we need to synthesise across datasets; saving data is essential for synthesis. Third, there has been a computational revolution in the types of analyses we can do and the amount of data that can be included [9]. This means that we can now finally perform powerful analyses of some of the exquisitely detailed data collected before the information revolution.

In recent years, there has also been a strong push from within scientific and scholarly communities for increased openness in science, including ecology and evolution (e.g., [10]). Calls for more transparency and accessibility in science are not new (e.g., [11]); the last decade, however, has seen a surge in general awareness and promotion of open science practises (e.g., open access publishing and open data, code, software, and peer-review) and their benefits [12]. These initiatives have not been without criticism, with many researchers unsure about sharing their data due to real or perceived concerns about data misuse and loss of control [13-15]. Others have acknowledged important caveats to the general appeal for openness (e.g., considerations about security, confidentiality, equity, and Indigenous data sovereignty and governance; [16-19]). Despite the legitimacy of (some of) these concerns, the benefits of data sharing are apparent [12,20]. Even so, large amounts of data remain private and unavailable for reuse. For example, in a sample of >4,000 ecology and evolution papers, only one in five papers (21.5%) had a data availability statement or associated open data [21], and less than half of archived datasets in ecology and evolution are reusable [21,22]).

Open science initiatives have developed rapidly, and the number of institutions, governments, funding agencies, and publishers who have implemented policies that require the open, permanent, and accessible sharing of data is increasing (e.g., FAIR data principles [23], the Ecological Society of America's new Open Research policy, the European Commission's OpenAIRE open access and open data policy). These requirements, and participation by scientists, will enhance our ability to evaluate, reuse, and synthesise increasingly rich and complex ecological data. However, open data policies are not retroactive and, therefore, do little to address issues of access to and preservation of previously-collected data [5]. Arguably, data collected prior to the adoption of widespread sharing practices remain a public good, funded by taxpayers and governments, so rescuing datasets to ensure their longevity and accessibility is imperative.

Here, we present general guidelines for implementing data rescue, with a focus on environmental data. These recommendations are based on past and ongoing data rescue projects by the Living Data Project, an initiative of the Canadian Institute of Ecology and Evolution

(CIEE), which aims to identify and secure vulnerable datasets and bring new life to them through collaborative analysis and synthesis (Box 2). We hope these guidelines will (a) focus attention on the current threats to the usability and integrity of previously-collected data, (b) stimulate broader consideration of the utility of previously-collected datasets for current research efforts, (c) encourage people with knowledge of unarchived data to preserve them, (d) provide a reference for those looking to apply data rescue techniques either *ad hoc* or as part of a broader initiative, and (e) help foster a strong culture of data stewardship such that data rescue becomes unnecessary in the future.

Guidelines for data rescue

Imperilled data can be found nearly everywhere (e.g., Box S1), such as non-profit organisations, conservation councils, academic institutions, and government agencies (think: historical data only available on paper records or digitised data stored only on floppy disks). Although data to be rescued are plentiful, discoverability is challenged by the very fact that they have not yet been rescued. Data rescue projects target data that are not properly archived, making them unfindable or inaccessible [23]. In ecology, for example, these issues lead to a low number of available datasets [21,24] and limit our capacity for knowledge synthesis. Ultimately, professional networks are valuable resources for finding languishing data hidden in field notebooks, file cabinets, old computers, and forgotten project files. As not all the data we need is research data [25], metadata, grey literature, and other auxiliary data may also be of importance. Additionally, movements for open data and transparency can help bring hidden data to light. Therefore, data rescue is embedded in a context of community building from the beginning to the data sharing step, in a feedback loop of outcomes: good sharing practices lead to more findable datasets.

Once data has been located, implementing a successful data rescue mission requires a strategic approach (Fig. 1 and Fig. 2). Some steps in data rescue are closely aligned with recommended practices in research data management. Several resources have already outlined "best" practises for data collection [21], management [22], and archiving [4,23,26,27], yet these are written with current or future data collection in mind and do not address historically-collected or unmanaged data. Below, we outline seven steps for data rescue, from identifying high-priority datasets to archiving and sharing them for (re)use.

1. Data prioritisation

Given potentially limited time (and money), data often needs to be prioritised for rescue over others. Prioritising data for rescue requires consideration along at least two axes: the scientific value of the data and the potential risk that the data will be lost (Fig. 1). In cases where data are of high value and at high risk, they should be given highest priority. Prioritisation becomes less obvious when data rank highly along just one of the axes of value and risk. In such instances, we suggest the focus should be on the value of the data, followed secondarily by risk (i.e., high value, low risk data should be prioritised over data that may be at high risk of loss but low value). The concepts of value and risk of loss are naturally subjective, and myriad factors (e.g., the interests of the rescuer or organisation, the combination of datasets to be compared) will impact how value and risk are assessed in each situation. As such, it is challenging to offer objectively clear guidelines for prioritisation. There are, however, general characteristics to consider when determining the value and risk of loss of a dataset.

High-value environmental datasets have some common features. Scale is a key factor, as datasets comprising long time series or a broad spatial extent are important for establishing temporal and spatial dynamics of change (e.g., population declines, range shifts, etc.). The age of a dataset may be relevant, as older datasets can establish important baselines for a species or system, and the value of such datasets increases with time. The subject of the data is also critical, as their societal value may be higher when involving species or ecosystems with conservation, cultural, or economic importance. Additional considerations include the rarity of the data (e.g., data from undersampled regions or ecosystems), uniqueness or irreplaceability (e.g., data from historical events, such as natural disasters), and the potential costs of recollection. Finally, potential future reuse is worth considering, with the highest value datasets having many, immediate potential use scenarios.

The risks of data loss are similarly multifold. Data can be physically lost, especially if there is only one copy (paper or digital). Data can be functionally lost when the datasets are unreadable due to defunct file formats (e.g., Lotus 1-2-3™) or obsolete storage media (e.g., floppy disks). Data can also be functionally lost when vital knowledge about collection or meaning is lost (e.g., because the collector/creator of the data is deceased, retired, or otherwise no longer active in their field). Ultimately, balancing the data's value and risk of loss is essential for effective prioritisation of data rescue efforts.

2. Team creation

Data rescue takes a team, with different roles needed at different points in the rescue process. We first consider those currently in possession of the data, who we collectively refer to as *data custodians*. These include:

(1) *data creators,* who are typically involved in generating the ideas that lead to the data's collection and retain the intellectual property rights and responsibilities for the data;

(2) *data collectors*, who generate or collect the original data and, therefore, provide valuable input for documenting the data; and

(3) *data stewards*, who are responsible for managing and maintaining the data (i.e., organising and keeping data archived, including instances where researchers have been bequeathed data or organisations act as guardians of data collected by past employees).

These roles are often played by the same person, though not always. For example, a graduate student may play all three roles as data creator, collector, and (temporary) steward, while the advisor may retain the data long-term as the principal investigator, thereby acting as data creator and (long-term) steward. Having at least one person who is a data creator, collector, or steward as part of the data rescue team is imperative for a successful data rescue mission.

A *data management expert* is another key role. Usually, a data manager plans the data lifecycle, but in a data rescue project this role is focused on organising and documenting the digitised datasets. This person will have the skills to connect datasets, clean and manage data, and compile previously unwritten information. Additionally, if any data are not in digital formats, a *data entry technician* will be an integral part of the team, ensuring all necessary data have been digitised in the appropriate format and validated against the original records.

3. Metadata creation

Metadata are information about the data, typically contained in a file separate from the dataset [31]. Metadata describe the data collection process (e.g., types of data collected, methodology, and contributors), variables in the dataset (e.g., column headings for tabular data; "data dictionary"), abbreviations, units of measurement, and other relevant information necessary to understanding how the data were generated and how to (re)use them (e.g., why some measurements are lacking; [27]). We recommend early creation of the metadata, as this often informs the remaining process and structure of the compiled dataset.

For datasets with more than one associated file, the metadata should also include a description of which data are contained in each file and how files are related. For datasets which include ongoing data collection, detailed metadata files are important to ensure that subsequently inputted data conform to existing standards and structure [32]. The metadata should be revised throughout the subsequent steps to incorporate details about the data rescue process (e.g., data manipulation, validation, or changes to database structure; Fig. 2).

Metadata file formats vary, often based on the type of data or chosen repository. In ecology, metadata are often provided in a "README" style text file that is, at a minimum, "human-readable" (i.e., a person can interpret the information contained in the file). Ideally, metadata should also be "machine-actionable", allowing computers to process and integrate datasets in an automated fashion (*Interoperability*) [23], enabling interaction with large volumes of data—a task that is not possible for humans to do.

A common format for creating metadata that are human- and machine-readable is a text file written in Extensible Markup Language (XML; for basic principles and examples, see https://www.xmlfiles.com/xml). A variation on XML called the Ecological Metadata Language (EML) is a set of suggested "tags" (variables) to create machine-actionable metadata in ecology [33,34](see https://eml.ecoinformatics.org/).

A recent alternative to XML is the use of schemas. For example, schema.org (https://schema.org) provides a collection of shared vocabularies to mark-up data in a standard fashion, allowing them to be understood by major search engines. The schema.org vocabulary is used in combination with a data-interchange language, such as JSON-LD, to structure and add information to data. Guidelines and examples of scientific use of schema.org are available from the Federation of Earth Science Information (https://wiki.esipfed.org/Main_Page) and Bioschemas (https://bioschemas.org). Tools also exist to help ecologists generate a schema and translate it to EML [35].

4. Data transfer and compilation

For effective collaboration, all team members should have access to the data and metadata files. However, this might only be possible if all files are already in a digital format;

any physical copies should first be photographed or scanned or entrusted to the team member responsible for data entry and validation. While the details of data compilation will need to be tailored to each dataset, the workflow should be as reproducible as possible. For example, any edits made to the data should be done in a file separate from the original; a digital file with untouched original data should always remain. All major decisions should be documented in the metadata.

In structuring the data, we recommend Wickham's [36] tidy data principles (also called "third normal form" relational data design [37]), which consist of 3 core concepts: (1) each variable has its own column, (2) each observation has its own row, and (3) each type of observational unit is in its own data table (e.g., individual-level measurements from a population, such as mass, in one table and population-level metrics, such as abundance, in another). If there are multiple data tables, they should be connected to each other by one or more variables that uniquely identify individual observations (i.e., primary keys in a relational database; [37]). While we advocate for tidy data principles, as they are most likely to generate a data structure that will be useful in subsequent analyses, sometimes alternative data structures will be preferred, such as site-by-species matrices for community-level data. Additionally, not all environmental data will be easily represented in tabular form, such as geospatial data or images, though other relevant standards may apply (see below). Finally, note that many data types are not well-suited to a relational database model and may benefit from other, equally valid frameworks (e.g., tree/graph-based data models in JSON).

5. Data cleaning and validation

Data cleaning consists of identifying and fixing issues and can be one of the most time-intensive steps. In addition to correcting typographical or entry errors, data cleaning includes checking for data completeness (i.e., all records are fully transcribed) and uniformity (i.e., variables and units are consistent). The International Organisation for Standardisation (ISO) provides standards for many common variables such as date-times (ISO 8601) and geographic coordinates (ISO 6709), and many tools exist to help with specific aspects of data cleaning (e.g., the *taxize* R package to check taxonomies; [38]).

Data validation involves the comparison of the dataset against a set of assertions. This is important for ensuring data quality and integrity by confirming that the structure and content of the data are appropriate. In data rescue, unlike most recently or currently collected data, data validation may come with the extra challenge that the original data custodians may be unreachable or deceased. As such, having as many original members of the data team as possible is particularly beneficial (Fig. 1, Step 2; see *Team creation*). Common data validation techniques include plotting the data to identify incorrect or improbable values, checking that the contents or dimensions of the data match expectations, cross-checking data from different columns or tables for mutual compatibility, and evaluating summary statistics or other outputs that characterise the data. In addition, many tools exist to help with the data validation process, including open-source, "point-and-click" software (e.g., OpenRefine) and programming tools (e.g., the *assertr* and *validate* R packages; [39,40]).

Although the exact data cleaning and validation steps will vary by dataset, many of the principles described in the *Data transfer and compilation* section are also relevant. Validation should be conducted as reproducibly as possible, and any errors should be corrected without manipulating the original (raw) files. Any changes should be well documented (e.g., as comments in the script or as notes in the metadata), as should the rationale behind the corrections.

Data custodians may also consider providing a checksum (e.g., md5) or cryptographic hash (e.g., SHA-256) foreach data file. Checksums and hashes are unique alpha-numeric signatures generated by an algorithm using the reference file as input information, such that even a trivial change in the contents or structure of the file will result in the production of a completely different output. A future potential user (including the original data creator) can then recalculate the hash upon accessing the archived data (see steps 6 and 7), compare it to the value stored in the metadata, and ensure data integrity prior to reuse.

6. Data archiving

Archiving data in non-proprietary formats is imperative for longevity and future accessibility. Non-proprietary formats are those which do not have a copyright or trademark and, therefore, are part of the public domain. Using non-proprietary formats ensures that anyone can access the data without needing specific software or in the event that the program becomes defunct. For example, tabular data should be stored in comma-separated values (.csv) format or text files (.txt) rather than proprietary formats such as Microsoft Excel® files (.xlsx). More recently, other open-source formats such as Apache parquet files (.parquet) have been developed, enabling highly efficient and compressed storage of “big” data. Unlike CSVs, parquet files also have the advantage of storing the schema (i.e., column/variable types; see *Metadata creation*) directly in the file metadata, reducing the chance that variables are incorrectly stored or used.

There is a growing movement to archive data on public data repositories rather than, or in addition to, private or institutional systems (e.g., lab hard drives). Many governments and funding agencies have recently implemented new data management protocols that encourage or mandate the archiving, though not necessarily sharing, of all data generated using their resources (see below; e.g., Canada’s Tri-agency Research Data Management Policy). Each year following publication, data that have not been publicly archived are 17% less likely to be recoverable [5] (see also [41]). As such, we consider public archiving to be an essential part of data rescue, since private archiving does not mitigate the possibility that data will need to be “re-rescued” in the future. Cleaned data and metadata should be placed in a repository, maintaining them in a secure and retrievable format. Importantly, the push for public archiving does not contradict the need for privacy or sensitivity associated with some datasets; it is possible to publicly archive data while maintaining restrictions on when and how the data are accessed. We suggest, however, that most environmental data should be openly accessible upon archiving, with some clear exceptions (e.g., data pertaining to threatened species or Indigenous data sovereignty; see below).

There are many data repositories from which to choose (see r3data.org for a comprehensive list), with some being generalised (e.g., Dryad, Dataverse, Figshare, Zenodo) and others more specified (e.g., DataONE for environmental data, GenBank for genetic sequences). Data repositories tend to use a distributed, decentralised approach to storing data and have contingency plans to ensure the longevity of archived datasets. Choice of repository will be influenced by whether the data will remain private or be made openly accessible upon upload, or soon thereafter [10]. Some repositories allow for the long-term storage regardless of whether data are made openly available (e.g., Dataverse), while others mandate open access (e.g., Dryad). Many archives also offer an option to place an embargo on the publication of data. Most data repositories establish a Digital Object Identifier (DOI), a unique identifier which remains constant for the lifetime of the object, even if the object or metadata change. For open data, we suggest explicitly stating the terms of use, such as whether authors should be contacted if the data are to be included in a publication, or adding a copyright statement, such as those from Creative Commons (e.g., CC0, CC-BY, etc.).

7. Data sharing

The final step in the data rescue workflow is ensuring the data meet open science standards. Open science principles include transparency, participation, and accessibility. These values can be addressed in different ways, sometimes making the process overwhelming for researchers who are not trained in data management. The FAIR and CARE principles, the first of which focuses on how data can be made useful and the second on how we can promote justice through responsibly sharing open data, summarise ways these values can be met through a combination of actions.

The **FAIR** principles aim to improve **F**indability, **A**ccessibility, **I**nteroperability and **R**eusability of datasets [23]. Providing human- and machine-readable metadata improves both the findability and accessibility of a dataset. Combined with proper archiving and identification, strong metadata helps increase the discoverability of datasets. As mentioned in the *Data archiving* section, adding a DOI makes the data trackable and citable. A comprehensive metadata file enables interoperability, or the ability of the data to be combined with other datasets in different ways and in different systems. Additionally, accessibility and reusability can be achieved through licences, which explicitly describe the usage and attribution rights of the data.

The **CARE** principles focus on datasets that used traditional knowledge or benefited somehow from Indigenous lands, promoting transparency and participation of open data [42; see also, the OCAP principles: https://fnigc.ca/ocap-training/]. They aim to address consideration of the **C**ollective benefit for Indigenous Peoples, **A**uthority to control (recognizing Indigenous data sovereignty), **R**esponsibility to be respectful with Indigenous Peoples involved in the dataset collection, and **E**thics (by assuring participation of Indigenous Peoples in the assessment of benefits, harms and usability of the data; [42]). These principles begin to address the larger, complicated history of colonialism in ecology, evolution, and related disciplines. While these guidelines were written with current and future data collection in mind, they are equally applicable to and important for previously collected data.

Carroll et al. (2021) provide valuable guidance on reconciling CARE and FAIR principles with Indigenous data-sovereignty at the forefront. Providing specific recommendations for addressing CARE principles in data rescue is challenging and beyond the scope of this paper; each project brings unique circumstances that are best navigated by the data custodians and Indigenous partners. In an ideal scenario, the data creator has established collaborations with relevant Indigenous communities, leading the data rescue effort to become another meaningful collaboration, collectively adjusting the data rescue workflow to address both FAIR and CARE principles—which, as Carroll et al., (2021) note, need not be in conflict. A full realisation of CARE principles would see Indigenous partners oversee data archiving and stewardship, with direct control over access to the repository [43]. Existing tools such as embargo periods (i.e., the delayed release of data) or controlled access (i.e., data hosted on a repository and available by request) may be useful in addressing concerns around sovereignty over sensitive data [13]. In cases where the data custodian has limited experience engaging with Indigenous communities, the potential to achieve CARE principles will depend upon the feasibility of developing trust and respectful relationships with the relevant Indigenous communities; given the devastating legacies of colonialism, this can take considerable time. Nevertheless, it would rarely be a misstep to request a meeting with local communities to communicate the goals of the data rescue project, highlighting the aim of achieving CARE principles in partnership with the community.

Conclusion

Ultimately, we hope to reach a point where data rescue is no longer needed. This requires researchers, funding agencies, and publishers to align their views around ethical and professional obligations to publicly archive data as well as a culture change that sees best practices in data managing, archiving, and sharing data become the default in publicly-funded research. To achieve this goal, data sharing and accessibility need to be prioritised as critical components of the scientific enterprise. First, there must be continued, long-term investment in data management [44]. Such investment includes not only infrastructure but also training and support for students and personnel [4,17]. Additionally, publishers, employers, and funding agencies must require accountability from researchers to preserve data in accessible formats and, if appropriate, make the data openly available[44]. Until these institutional-level paradigm shifts occur, smaller-scale and innovative data rescue is integral to environmental data curation.

Currently, training in data management and shifting regulations regarding data availability have focused on present and future data. With such a strong eye to the future, however, data of the past is being left behind. Data rescue presents an opportunity to mitigate this loss of historical data while also providing additional, less tangible benefits. In the CIEE Living Data Project, our mission of breathing life into languishing data is concomitant with training the next generations of scientists in data management best practises and forging connections amongst researchers across a wide variety of career stages and trajectories, thus ensuring the longevity of scientific knowledge and preparing students for a data-rich future.

Acknowledgements

The Living Data Project (LDP) is a collaborative initiative by researchers at institutions across Canada: University of British Columbia-Vancouver, University of British Columbia-Okanagan, University of Regina, McGill University, and Université de Montréal. The authors acknowledge that we live and work on the traditional, ancestral, treaty, and unceded territories of many Indigenous peoples, including the Coast Salish Peoples, xwməθkwəy̓əm (Musqueam), Syilx (Okanagan), nêhiyawak (Cree), anihšināpēk (Saulteaux), Dakota, Lakota, Nakoda, Attawanderon, Mississaugas, kanien’kehà:ka (Mohawk), and Haudenosaunee, and the homeland of the Métis/Michif Nation.

We thank LDP members and partner organisations for thoughtful discussions on data rescue, best practices in data management, and fostering more open and transparent science, especially those involved in Box 2: Jenna Loesberg, Amelia Hesketh, Ellen Macdonald, Justine Karst, Andrea Brown, and Harold Eyster.

The LDP is funded by a Collaborative Research and Training Experience (CREATE) grant to the Canadian Institute of Ecology and Evolution from the Natural Sciences and Engineering Research Council of Canada. EKB, JBB, and GTH were supported by the CREATE grant; EKB was also supported by the University of Regina. DGR was supported by the European Union’s Horizon 2020 research and innovation program, under the Marie Skłodowska-Curie grant agreement (No. 838237-OPTIMISE).

Figure 1. Prioritising data for rescue: balancing the value of the data and its risk of loss.

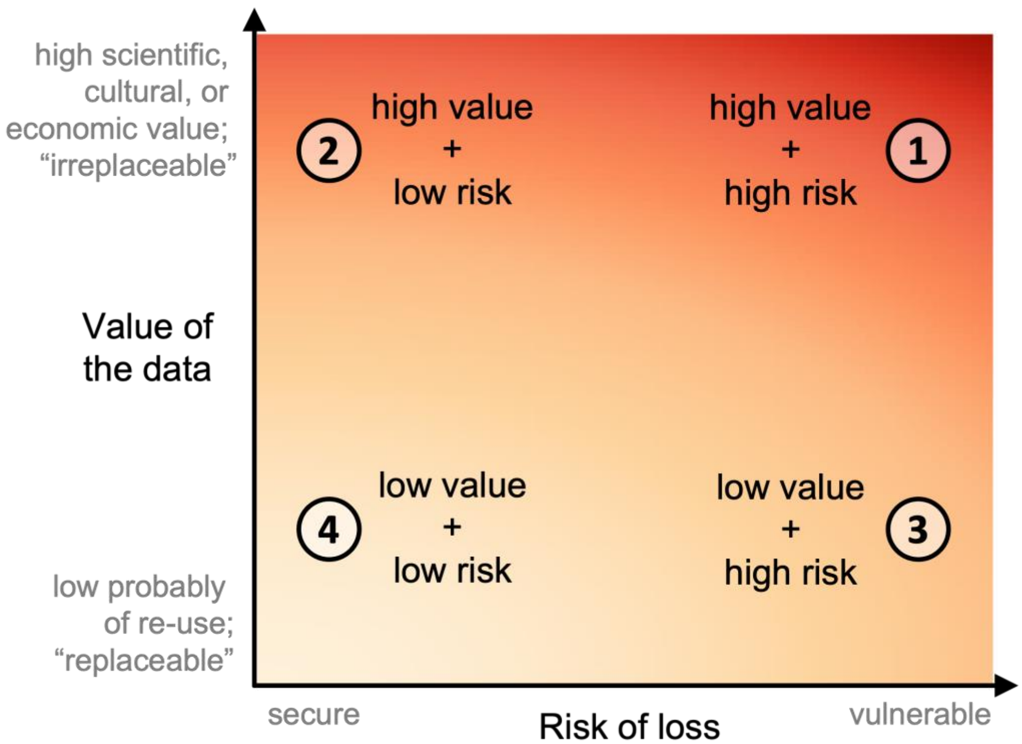


Figure 1. Prioritising data for rescue: balancing the value of the data and its risk of loss. With many datasets in need of preservation and limited resources, the first step in the data rescue process requires developing a list of priorities for consideration and identifying relevant datasets (Fig. 2). We consider data prioritisation to be a balance between the assessed value of a dataset in question and the potential risk of its loss in the absence of intervention (see *Data prioritisation* under *Guidelines*).

[Alt text: Figure 1 shows a two-dimensional colour gradient to help conceptualise one approach to data prioritisation. "Risk of loss" is on the horizontal axis, with the left-hand side labelled 'secure' and right-hand side 'vulnerable'. "Value of the data" is on the vertical axis, with the bottom labelled 'low probability of reuse; replaceable' and the top labelled 'high scientific, cultural, or economic value; irreplaceable'. The plot area ranges from red in the top right ("1. high value + high risk"), to reddish orange in the top left ("2. high value + low risk"), to orangey-yellow in the bottom right ("3. low value + high risk"), to yellowish white in the bottom left (“4. low value + low risk”).]

Figure 2. Steps in the data rescue assembly line.

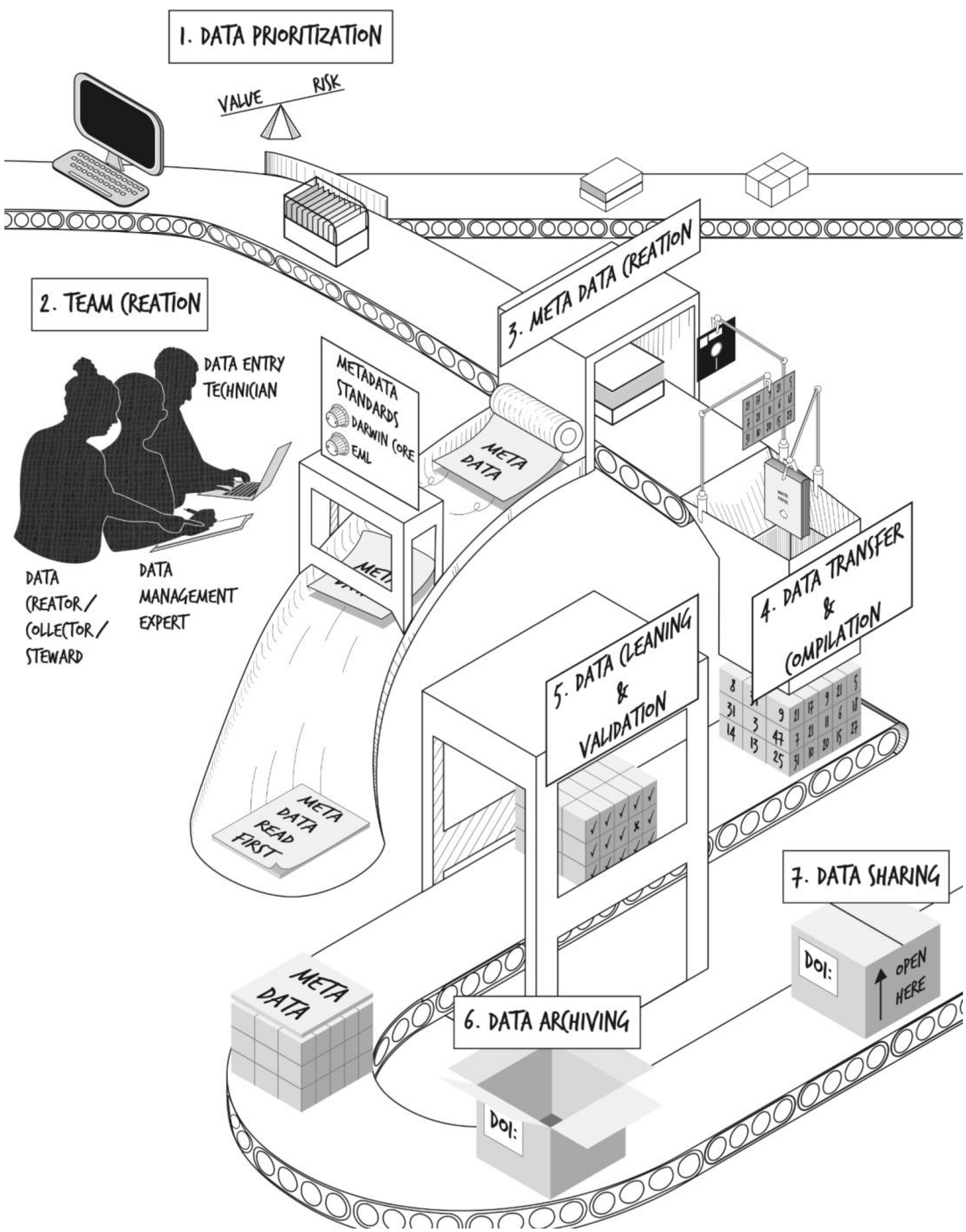


Figure 2. Steps in the data rescue assembly line. First, data must be prioritised for rescue (Step 1). After team creation (Step 2) and metadata creation (Step 3), the data must be transferred and compiled into a logical format (Step 4). After data cleaning and validation (Step 5) is complete, the finalised data and metadata should be archived on a long-term data repository (Step 6). The ultimate goal is to have the rescued data openly available for reuse (Step 7).

Box 1. Spilt oil, spent money, and lost data

In 1989, the oil tanker *Exxon Valdez* struck the Bligh Reef in Prince William Sound, less than 2.5 km from the Alaskan shore. As a result, approximately 37,000 tonnes of crude oil spilled into the sound, leading to catastrophic short- and long-term ecological consequences. The *Exxon Valdez* Oil Spill Trustee Council (EVOSTC) was established in 1991 to oversee the spending of funds from a civil settlement in 1991 between *Exxon*, the United States federal government and the state government of Alaska. A large portion of funds were directed towards determining and monitoring the impacts of the oil spill on oceanographic, environmental, and ecological conditions. Prior to 2003, there was no requirement for data preservation or availability; afterwards, all projects were awarded under explicit conditions from EVOSTC that data be preserved and made publicly available [45]. In their annual report from 2010, the EVOSTC notes that some $151.2 million USD were spent on "research, monitoring, and general restoration" during 1992-2010 fiscal years [46].

From 2012-2014, a group of researchers from the National Center for Ecological Analysis & Synthesis (NCEAS) worked to recover the historical datasets funded by EVOSTC, focusing specifically on data collected between 1989-2010 [45]. Of the 419 projects funded by EVOSTC during this time, only 27% of the datasets were able to be recovered; after a total of 5 years hunting down datasets, this grew to 30% [45].

Using these numbers, we can roughly estimate the money spent on research for which the data are unrecoverable (70% of datasets): ~**$105 million USD was spent collecting data that are no longer recoverable and, therefore, effectively non-existent to science.** While we do not know the distribution of years from which data were recovered or how money was allocated by year, this is likely a conservative estimate given that the original cost does not include the first 3 years following the spill, when extensive ecological assessments would have been completed.

Box 2. Data Rescue Examples from the Living Data Project.

Seeing the Forest Data for the Trees

Upon the retirement or death of a professor, students or colleagues sometimes must take the reins and piece together documents and data from decades-old research projects.

Step 1: Data prioritisation

Dr. George H. La Roi was a professor of forest ecology at the University of Alberta (UofA) for 35 years. Upon his passing, La Roi's children bequeathed his legacy of highly valuable data to his former colleague who had earlier taken over sampling some of his long-term plots. With no living data creator and the data in unorganised boxes containing unsorted datasheets, documents, CD-ROMs, and picture slides (Box 2.1), the data was at high risk of loss.

Step 2: Team creation

Two of Dr. La Roi's colleagues served as data stewards. Two graduate interns worked as data management experts, along with several undergraduate data entry technicians who sorted, entered, and digitised the data.

Step 3: Metadata creation

Thankfully, one of the loose files was a report with methodology for many of the data collection events. Initially, inventory on the data needed to be done. Finalised metadata were written and consolidated into one document for future reuse; while most of the data had clear documentation, some data were lost due to undetermined variable definitions and units.

Step 4: Data transfer and compilation

The boxes of data were sent to the graduate students, and digitised data was transferred via a cloud-based service. The interns recovered data recorded at two different locations, both of which included similar measurements from plants. Some data were stored as printed scans of hand-filled datasheets, and thus required digitisation. Other data, which had already been entered and digitised, were stored in hundreds of text files which required extensive reformatting before they could be compiled into tidy, usable datasets.

Step 5: Data cleaning and validation

Standard data cleaning and validation procedures were conducted, such as removing character values in numeric columns, checking the data for obvious outliers, etc. Extensive work was done to ensure consistent taxonomy throughout the decades of data collection.

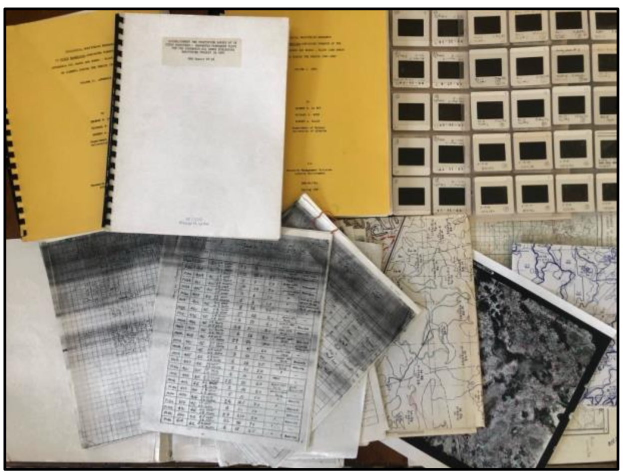
Step 6: Data archiving

The data and metadata of this expansive dataset has been archived and made publicly available through UofA's Dataverse repository [47] with a CC-BY licence.

Step 7: Data sharing

All files associated with the data follow FAIR data guidelines, with extensive metadata, files in non-proprietary file formats, and uploaded to an open data repository with a DOI.

Box 2.1. *Photograph of loose data sheets, maps, reports, and picture slides; these items and many more filled the boxes of research material left behind by Dr. La Roi. Image credit: A. Hesketh.*



Out of the Archives and into the (Digital) Light of Day

Theses and dissertations of former graduate students represent a rich source of historical data. In particular, those prepared prior to the advent of modern computer technologies and software (e.g., word processors) may contain troves of raw and summary data that remain un-digitized.

Step 1: Data prioritisation

This project was focused on securing the data contained in three, historical graduate theses from the University of British Columbia (UBC). While the specific questions and research topics differed, all three surveyed bird abundances in the same (or nearby) sites in Greater Vancouver, British Columbia, and combined present an opportunity to establish a baseline against which to compare current and future trends (Box. 2.S1). These data were prioritised because they were both at-risk (much of the data existed only in non-digital formats and none of the datasets are in active use) and deemed of high value (the data provide a valuable frame of reference for studying changes in urban bird diversity).

Step 2: Team creation

The project was proposed by a graduate student at UBC and was carried out in collaboration with a data rescue intern. As with the previous case, the original data creators were not directly involved in the data rescue, although one individual did provide a digital copy of the data contained within their thesis.

Step 3: Metadata creation

Given the extensive data manipulation required, clear metadata were developed to document the various steps taken to generate the final datasets and document other details from the theses that were not captured during the digitization process.

Step 4: Data transfer and compilation

The intern first worked to transcribe and digitise the data from the two earlier theses, which were only available from the thesis repository as scans of typewritten documents. Among other challenges, digitisation required the conversion of non-standard data types (Box 2.2) into "tidy" forms that could be interpreted programmatically. Data from the third thesis [50] were made available by the original author in a spreadsheet and so only required cleaning, manipulation, and conversion to a non-proprietary format.

Step 5: Data cleaning and validation

Later work included efforts to rationalise the datasets so they might be used in combination with each other (e.g., standardising column names and combining similar tables into a single file).

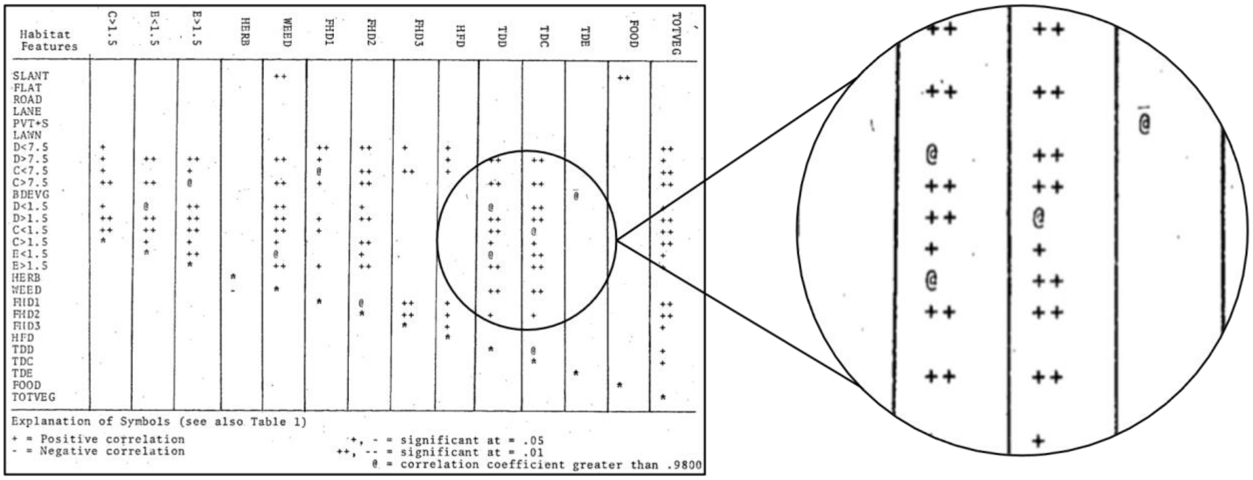
Step 6: Data archiving

The data have been archived on the UBC Scholars Portal Dataverse repository [48-50] and cross-linked to the original theses.

Step 6: Data sharing

The datasets have been archived following FAIR principles, include detailed metadata describing the data rescue process, use non-proprietary file formats, and have permanent DOIs.

Box 2.2. *Example of non-standard data to be rationalised and digitised, representing the significance of correlations between habitat features. These symbols were converted to numeric factors during digitization. Reproduced with modification from Lancaster [49] (see: Appendix 4, p. 103-104 therein).*



Habitat Features	C>1.5	E<1.5	E>1.5	HERB	WEED	FHD1	FHD2	FHD3	HFD	TDD	TDC	TDE	FOOD	TOTVEG
SLANT					++								++	
FLAT														
ROAD														
LANE														
PVT+S														
LAWN														
D<7.5	+					++	++	+	+					++
D>7.5	+	++	++		++	+			+	++	++			+
C<7.5	+		+			@	++	++	+					++
C>7.5	++	++	@		++	+	++			++	++			++
BDEVG												@		
D<1.5	+	@	++		++		+			@	++			
D>1.5	++	++	++		++	+	++			++	++			++
C<1.5	++	++	++		++	+				++	@			++
C>1.5	*	+	+		+		++			+	+			++
E<1.5		*	++		@		+			@	++			+
E>1.5			*		++	+	++			++	++			
HERB				*										
WEED				-	*					++	++			
FHD1						*	@	++	+					++
FHD2							*	++	+	+	+			++
FHD3								*	+					+
HFD									*					
TDD										*	@			+
TDC											*			+
TDE												*		
FOOD													*	
TOTVEG														*

Explanation of Symbols (see also Table 1)

+ = Positive correlation
- = Negative correlation

+, - = significant at = .05
++, -- = significant at = .01
@ = correlation coefficient greater than .9800

References

1. McClenachan L, Ferretti F, Baum JK. 2012 From archives to conservation: why historical data are needed to set baselines for marine animals and ecosystems. *Conserv. Lett.* **5**, 349-359. (doi:10.1111/j.1755-263X.2012.00253.x)
2. Gatti G, Bianchi CN, Parravicini V, Rovere A, Peirano A, Montefalcone M, Massa F, Morri C. 2015 Ecological change, sliding baselines and the importance of historical data: lessons from combining observational and quantitative data on a temperate reef over 70 years. PLoS One **10**, e0123268. (doi:10.1371/journal.pone.0118581)
3. Willis KJ, Araùjo MB, Bennett KD, Figueroa-Rangel B, Freud CA, Myers N. 2007 How can a knowledge of the past help to conserve the future? Biodiversity conservation and the relevance of long-term ecological data. *Phil. Trans. R. Soc. Lond. B* **362**, 175-187. (doi:10.1098/rstb.2006.1977)
4. Renaut S, Budden AE, Gravel D, Poisot T, Peres-Neto P. 2018 Management, archiving, and sharing for biologists and the role of research institutions in the technology-oriented age. *Bioscience* 68, 400-411. (10.1093/biosci/biy038)
5. Vines TH, Albert AYK, Andrew RL, Débarre F, Bock DG, Franklin MT, Gilbert KJ, Moore J, Renault S, Rennison DJ. (2014). The availability of research data declines rapidly with article age. *Curr. Biol.* **24**, 94-97. (doi:10.1016/j.cub.2013.11.014)
6. Buxton RT, Nyboer EA, Pigeon KE, Raby GD, Rytwinski T, Gallagher AJ, Schuster R, Lin H-Y, Fahrig L, Bennett JR, et al. 2021 Avoiding wasted research resources in conservation science. *Converv. Sci. Pract.* **3**, e329. (doi:10.1111/csp2.329)
7. Srivastava, DS, JL McCune, HK Lotze. 2017. Environmental change: a historical perspective In: Reflections on Canada Illuminating our Biggest Possibilities and Challenges at 150 Years (Editor: Philippe Tortell). Peter Wall Institute for Advanced Studies, Vancouver Canada. (ISBN:978-0888652676)
8. Price MH, Connors BM, Candy JR, McIntosh B, Beacham TD, Moore JW, Reynolds JD (2019). Genetics of century-old fish scales reveal population patterns of decline. *Conserv. Lett.* **12**, e12669. (doi:10.1111/conl.12669)

9. Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, Batcheller AL, Duke CS, Porter JH. 2013. Big data and the future of ecology. *Front. Ecol. Environ.* **11**, 156-62. (doi:10.1890/120103)
10. O'Dea RE, Parker TH, Chee YE, Culina A, Drobniak SM, Duncan DH, Fidler F, Gould E, Ihle M, Kelly CD. 2021 Towards open, reliable, & transparent ecology and evolutionary biology. *BMC Biol.* **19**, 1-5. (doi:10.1186/s12915-021-01006-3)
11. Eamon W. 1985 From the secrets of nature to public knowledge: the origins of the concept of openness in science. *Minerva* **23**, 321-347. (doi:10.1007/BF01096442)
12. Powers SM, Hampton SE. 2019 Open science, reproducibility, and transparency in ecology. *Ecol Appl.* **29**, e01822. (doi:10.1002/eap.1822)
13. Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, Cain KE, Kokko H, Jennions MD, Kruuk LE. 2014 Troubleshooting public data archiving: suggestions to increase participation. *PLoS Biol*. **12**, e1001779. (doi:10.1371/journal.pbio.1001779)
14. Smith R, Roberts I. 2016 Time for sharing data to become routine: the seven excuses for not doing so are all invalid. *F1000 Res.* **5**, 781. (doi:10.12688/f1000research.8422.1)
15. Stieglitz S, Wilms K, Mirbabaie M, Hofeditz L, Brenger B, López A, Rehwald S. 2020 When are researchers willing to share their data? - Impacts of values and uncertainty on open data in academia. *PLoS One* **15**, e0234172. (doi:10.1371/journal.pone.0234172)
16. Borgman CL. 2018 Open data, grey data, and stewardship: universities at the privacy frontier. *Berkeley Tech. Law J.* **33**, 365-412. (doi:10.15779/Z38B56D489)
17. Walter M, Suina M. 2018 Indigenous data, indigenous methodologies and indigenous data sovereignty. *Int. J. Soc. Res. Methodol.* **22**, 233-243. (doi:10.1080/13645579.2018.1531228)
18. Lennox RJ, Harcourt R, Bennett JR, Davies A, Ford AT, Frey RM, Hayward MW, Hussey ME, Iverson SJ, Kays R, et al. (2020). A novel framework to protect animal data in a world of biosurveillance. *BioScience* **70**, 468-476. (doi:10.1093/biosci/biaa035)
19. Buck S. 2021 Beware performative reproducibility. *Nature* **595**, 151. (doi:10.1038/d41586-021-01824-z)
20. Soeharjono S, Roche DR. 2021 Reported individual costs and benefits of sharing open data among Canadian academic faculty in ecology and evolution. *BioScience* **71**, biab024. (doi:10.1093/biosci/biab024)

21. Roche DG, Berberi I, Dhane F, Lauzon F, Soeharjono S, Dakin R, Binning SA. 2022 Slow improvements to the archiving quality of open datasets in evolution and ecology. *Proc. R. Soc. Lond. B* (in press). (doi:10.32942/osf.io/d63js)

22. Roche DG, Kruuk LEB, Lanfear R, Binning SA. 2015 Public data archiving in ecology and evolution: how well are we doing? *PLoS Biol.* **13**, e1002295. (doi:10.1371/journal.pbio.1002295)

23. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton, G., Axton, M., Baak, A., Blomberg N, Boiten J-W, Bonino da Silva Santos L, Bourne PE, et al. 2016 The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018. (doi:10.1038/sdata.2016.18)

24. Roche DG, Raby GD, Norin T, Ern R, Scheuffele H, Skeeles M, ... Binning SA. 2022. Paths towards greater consensus building in experimental biology. *J. Exp. Biol.* 225: jeb243559. (doi:10.1242/jeb.243559)

25. Gregory KM, Groth P, Scharnhorst A, Wyatt S. 2020. Lost or found? Discovering data needed for research. *Harvard Data Sci. Rev.*, **2**, e38165eb. (doi:10.1162/99608f92.e38165eb)

26. Broman KW, Woo KH. 2018 Data organization in spreadsheets. *Am. Stat.* **72**, 2-10. (doi:10.1080/00031305.2017.1375989)

27. BES. 2018 A guide to data management in ecology and evolution. BES Guides to Better Science. British Ecological Society, London, UK. (https://www.britishecologicalsociety.org/wp-content/uploads/2019/06/BES-Guide-Data-Management-2019.pdf)

28. Cook RB, Olson RJ, Kanciruk P, Hook LA. 2001 Best practices for preparing ecological data sets to share and archive. *Bull. Ecol. Soc. Am.* **82**, 138-141. (https://www.jstor.org/stable/20168543)

29. White EP, Baldridge E, Brym ZT, Locey KJ, McGlinn DJ, Supp SR. (2013). Nine simple ways to make it easier to (re) use your data. *Ideas in Ecology and Evolution*, 6, 1–10. (doi:10.4033/iee.2013.6b.6.f)

30. Whitlock MC. 2011 Data archiving in ecology and evolution: best practices. *Trends Ecol. Evol.* **26**, 61-65. (doi:10.1016/j.tree.2010.11.006)

31. Michener WK, Brunt JW, Helly JJ, Kirchner TB, Stafford SG. 1997 Non-geospatial metadata for the ecological sciences. *Ecol. Appl.* **7**, 330-342. (doi:10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2)

32. Yenni GM, Christensen EM, Bledsoe EK, Supp SR, Diaz RM, White EP, Ernest SM. 2019 Developing a modern data workflow for regularly updated data. *PLoS Biol.* **17**, e3000125. (doi:10.1371/journal.pbio.3000125)

33. Fegraus EH, Andelman S, Jones MB, Schildhauer M. 2005 Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bull. Ecol. Soc. Am.* **86**, 158-68. (http://www.jstor.org/stable/bullecosociamer.86.3.158)

34. Jones MB, O'Brien M, Mecum B, Boettiger C, Schildhauer M, Maier M, Whiteaker T, Earl S, Chong S. 2019 Ecological Metadata Language version 2.2.0. KNB Data Repository. (doi:10.5063/F11834T2)

35. Boettiger C. 2019. Ecological metadata as linked data. *J. Open Source Softw.* **4**, 1276. (doi:10.21105/joss.01276)

36. Wickham H. 2014 Tidy Data. *J. Stat. Softw.* **59**, 1-23. (doi:10.18637/jss.v059.i10)

37. Codd EF. 1990 *The Relational Model for Database Management: Version 2.* Addison-Wesley Longman Publishing. (ISBN:978-0-201-14192-4)

38. Chamberlain S, Szocs E. 2013 taxize – taxonomic search and retrieval in R. *F1000 Res.* **2**, 191. (doi:10.12688/f1000research.2-191.v2)

39. Fischetti T. 2020 assertr: assertive programming for R analysis pipelines. R package version 2.7. (https://CRAN.R-project.org/package=assertr)

40. van der Loo MPJ, de Jonge E. 2021 Data validation infrastructure for R. J. Stat. Softw. 97, 1–31. (doi:10.18637/jss.v097.i10)

41. Tedersoo L, Küngas R, Oras E, Köster K, Eenmaa H, Leijen Ä, Pedaste M, Raju M, Astapova A, Lukner H, Korgerman K, Sepp T. 2021 Data sharing practices and data availability upon request differ across scientific disciplines. *Sci. Data* **8**, 192. (doi:10.1093/biosci/biab024)

42. Carroll SR, Garba I, Figueroa-Rodríguez OL, Holbrook J, Lovett R, Materechera S, Parsons M, Raseroka K, Rodriguez-Lonebear D, Rowe R, et al. 2020 The CARE principles for Indigenous data governance. *Data Sci. J.* **19**, 43. (doi:10.5334/dsj-2020-043)

43. Carroll SR, Herczog E, Hudson M, Russell K, Stall S. 2021. Operationalizing the CARE and FAIR Principles for Indigenous data futures. Sci Data 8, 108. (doi:10.1038/s41597-021-00892-0)

44. Mons B. 2020 Invest 5% of research funds in ensuring data are reusable. *Nature* **578**, 491. (doi:10.1038/d41586-020-00505-7)

45. Jones MB, Blake R, Couture J, Ward C. 2018 Collaborative data management and holistic synthesis of impacts and recovery status associated with the *Exxon Valdez* oil spill. *Exxon Valdez Oil Spill Long-Term Monitoring Program (Gulf Watch Alaska) Final Report* (project 16120120). *Exxon Valdez* Oil Spill Trustee Council, Anchorage, Alaska. (http://www.gulfwatchalaska.org/wp-content/uploads/2018/08/16120120-Jones-et-al.-2018-Final-Report.pdf)

46. EVOSTC. 2012 2010 Annual Report. *Exxon Valdez Oil Spill Trustee Council*. (https://evostc.state.ak.us/media/4411/2010annualreport.pdf)

47. Hesketh A, Loesberg J, Bledsoe EK, Karst J, Macdonald E. 2021. Seasonal and annual dynamics of western Canadian boreal forest plant communities: a legacy dataset spanning four decades. Scholars Portal Dataverse, V1 (doi: 10.5683/SP3/PZCAVE)

48. Brown A, Eyster H, Lancaster RK. 2021a Data for: Bird communities in relation to the structure of urban habitats. *Scholars Portal Dataverse*. (doi:10.5683/SP2/YD6N7C)

49. Brown A, Eyster H, Melles SJ. 2021b Data for: Effects of landscape and local habitat features on bird communities: a study of an urban gradient in greater Vancouver. *Scholars Portal Dataverse*. (doi:10.5683/SP2/BPLPAP)

50. Brown A, Eyster H, Weber WC. 2021c Data for: Birds in cities: a study of populations, foraging ecology and nest-sites of urban birds. *Scholars Dataverse Portal*. (doi:10.5683/SP2/K5LMLA)