Data rescue: saving environmental data from extinction

Ellen K. Bledsoe [1,2,†,*] (0000-0002-3629-7235), Joseph B. Burant[1,3,4,†,*] (0000-0002-0713-3100), Gracielle T. Higino[1,5,†] (0000-0003-2791-8383), Dominique G. Roche[1,6] (0000-0002-3326-864X), Sandra A. Binning[1,4] (0000-0002-2804-9979), Kerri Finlay[1,2] (0000-0001-6835-8832), Jason Pither[1,7] (0000-0002-7490-6839), Laura S. Pollock[1,3] (0000-0002-6004-4027), Jennifer M. Sunday[1,3] (0000-0001-9372-040X), Diane S. Srivastava[1,6,*] (0000-0003-4541-5595)

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.

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

Running Headline

Data rescue: saving environmental data

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 on 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 prioritization 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 research, open data, open science, reproducibility

Total word count (excluding title page and supplementary materials) = 7,821

- Abstract = 199
- Main text (excluding abstract, figure captions, boxes, and references) = 4,037
- Figures: 2
  - Figure 1 caption = 80
  - Figure 2 caption = 83
- Boxes: 3
  - Box 1 = 458 words
  - Box 2 = 408 words, 2 images
  - Box 3 = 1,202 words, 4 images
- References = 49 (1,354 words)

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"; McClenachan et al., 2012), understanding how species and communities respond to environmental change (Gatti et al., 2015), and designing and evaluating the outcomes of management efforts (Hawkin et al., 2013; Willis et al., 2007). 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 (Renaut et al., 2018). Yet, despite their high value, data are often misplaced, filed away, or otherwise rendered unusable, often through poor data management practices (Vines et al., 2014). In their unusable and "at-risk" state, these data represent an egregious waste of resources expended on their collection (Buxton et al., 2021; Box 1). 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 realize 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.

In recent years, there has been a strong push from within the scientific and scholarly communities for increased openness in the practice of science, including in ecology and evolution (e.g., O'Dea et al., 2021). Calls for more transparency and accessibility in science are not new (e.g., Eamon, 1985); the last decade, however, has seen a surge in general awareness and promotion of open science practices (e.g., open access publishing and open data, code, software, and peer-review) and their benefits (Powers & Hampton, 2019). 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 (Roche et al., 2014; Smith & Roberts, 2016; Stieglitz et al., 2020). Others have acknowledged important caveats to the general appeal for openness (e.g., valid considerations about security, confidentiality, equity, and Indigenous data sovereignty and governance; Borgman, 2018; Walter & Suina, 2018; Lennox et al., 2020; Buck, 2021). Despite the legitimacy of (some of) these concerns, the benefits of data sharing are apparent (Powers & Hampton; 2019; Soeharjono & Roche, 2021). Even so, large amounts of data remain private and unavailable for reuse by other scientists. For example, in a sample of more than 4,000 ecology and evolution papers, only one in five papers (21.5%) had a data availability statement or associated open data (Roche et al., 2021), and less than half of archived datasets in ecology and evolution are reusable (Roche et al., 2015; Roche et al., 2021).

Open science initiatives have developed rapidly, and the last few years have seen a rise in the number of institutions, governments, funding agencies, and publishers who have implemented policies that require the open, permanent, and accessible sharing of data (e.g., FAIR data principles [see *Data sharing*; Wilkinson et al., 2016], the Ecological Society of America's new Open Research policy, and 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 synthesize 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 (Vines et al., 2014). 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. We include examples using historical (Box 2) and recent data rescue efforts (Box 3). We anticipate 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 access to or knowledge of unarchived data to work towards their preservation, (d) provide a reference for those looking to apply data rescue techniques in the context of their own work, and (e) help foster a strong culture of data stewardship such that data rescue becomes unnecessary in the future.

Guidelines for data rescue

Imperiled data can be found nearly everywhere, such as non-profit organizations, conservation councils, academic institutions, and government agencies (think: historical data only available on paper records in basement filing cabinets, digitized data stored only on floppy disks, etc.). Finding data to rescue is usually the easy part; implementing a successful data rescue mission, however, requires a more strategic approach (Fig. 1). Some steps involved in data rescue are closely aligned with recommended practices in research data management (see *Metadata*, *Data Compilation*, *Validation*, *Archiving* and *Sharing* sections). Several resources have already outlined "best" practices for data collection (Broman & Woo, 2018), management (e.g., BES, 2018), and archiving (Cook et al., 2001; Renault et al., 2018; Whitlock, 2011; White et al., 2013), 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 key steps for data rescue, from identifying high-priority datasets to archiving and sharing them for (re)use.

1. Data prioritization

Prioritizing data for rescue requires consideration of both the scientific value of the data and the potential risk that the data will be lost (Fig. 2). Data of high value and at high risk should be given highest priority, while data which rank highly along just one of the axes of value and risk should be considered moderate priorities. The concepts of value and risk of loss are naturally subjective, but there are some general factors to consider when determining these characteristics of a dataset.

High-value environmental datasets have some common features. Scale is a key factor, as datasets comprising long time series or covering a broad spatial extent are often 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 the societal value of the data may be higher when it involves species or ecosystems with conservation, cultural, or economic value. Additional considerations are the rarity of the data (e.g., data from an undersampled region or ecosystem), their uniqueness or irreplaceability (e.g., data from a historical event, such as a natural disaster), and the potential costs of recollecting the data, if that is a possibility. Finally, how the data might be re-used in the future is important, with the most high-value datasets having many, immediate potential use scenarios. This is, perhaps, the most difficult (and subjective) factor to assess.

The risks of data loss are similarly multifold. Data can be physically lost, and this risk is highest for datasets for which there is only one copy (paper or digital). Data can also be functionally lost when the datasets are unreadable because they are in older or defunct file formats (e.g., Lotus 1-2-3™) or in obsolete storage media (e.g., floppy disks). Data can also be functionally lost when vital knowledge about collection or meaning of the data 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 prioritization 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: *data creators* 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, even if not directly involved in collecting or managing the data products; *data collectors* generate or collect the original data and, therefore, provide valuable input for documenting the data (see *Metadata creation*); and *data stewards* are responsible for managing and maintaining the data (i.e., organizing and keeping data safely archived, including instances where researchers have been bequeathed data or organizations act as custodians of data collected by past employees). These roles are often played by the same person, though not always. For example, in a mentee-mentor relationship such as that between a graduate student and supervisor, the 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 is the one that plans the data lifecycle, but in a data rescue project this role is mainly focused on organizing and documenting the digitized datasets. This person will have the skills to connect datasets, clean and manage data, and compile previously unwritten information. Additionally, if any data have not been entered into a digital format, a *data entry technician* will be an integral part of the team, ensuring that all necessary data have been digitized 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 (Michener et al., 1997). Metadata generally describe the data collection process (e.g., types of data collected, methodology, and contributors), a description of the 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; British Ecological Society, 2018). We recommend early creation of the metadata, as this often informs the remaining data rescue process and structure of the compiled dataset.

For datasets with more than one associated file, the metadata should also include a description of the database structure, which data are contained in each file, and how files or tables relate to each other. For datasets which include ongoing data collection, detailed metadata files are important to ensure that subsequent data added to the database conform to the appropriate standards and existing structure (Yenni et al., 2019). The metadata will likely need to be revised after *Data compilation* (Step 5) and before *Data archiving* (Step 6) to incorporate details about the data rescue process (e.g., data manipulation, validation, or changes to the structure of the dataset or database; Fig. 1).

The metadata file format varies (often dependent on the type of data or chosen repository), but one useful format is a text file written in Extensible Markup Language (XML; see examples at https://www.xmlfiles.com/xml/). Tools like XML have been developed specifically for writing and storing metadata in a format that is both human *and* machine readable, not only ensuring that end users understand the data structure but also facilitating use by other software/programming tools (e.g., search engines) that may rely on metadata being available in a standardized form. Each variable is stored as a "tag," and its description is stored between tags. There is a variation of XML called Ecological Metadata Language (EML; Fegraus et al., 2005; Jones et al., 2019; see https://eml.ecoinformatics.org/) which offers a set of suggested tags specific to describing environmental data.

4. Data transfer and compilation

For the data rescue team to work effectively, 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 either be photographed or scanned first or entrusted to the team member responsible for data entry and validation. From there, discussion about how the data should be compiled most effectively can ensue. 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. Additionally, all major decisions should be documented in the metadata.

In structuring the data, we recommend following Wickham's (2014) tidy data principles, 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; Codd, 1990). 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 compromises will need to be made.

5. Data cleaning and validation

Following data entry and compilation, data cleaning can be one of the most time-intensive steps of the data management process. Data cleaning is the process of identifying and fixing issues, such as data entry errors or incomplete records. In addition to common steps like correcting typographical or entry errors, data cleaning commonly includes checking for data completeness (i.e., that the data from all records are fully and correctly transcribed) and uniformity (i.e., that variables are recorded in a consistent way for all records, ensuring common measurement units, etc.) and otherwise ensuring the data conform to expected standards. For ecological or biodiversity data, other common data cleaning steps include checking for common date formats (e.g., the International Organization for Standardization (ISO) 8601 standard recommends date-time objects be recorded as YYYY-MM-DD hh:mm:ss + UTC offset), ensuring geographic coordinates are complete and standardized (e.g., ISO 6709 applies to the representation of spatial information), and correcting misspellings or synonyms in taxonomic information. Many tools have been developed to help with specific aspects of data cleaning (e.g., the *taxize* package in R can be used to correct taxonomies; Chamberlin & Szocs, 2013).

Related to data cleaning, data validation involves the comparison of the dataset against a set of assertions determined *a priori* (e.g., dry body mass of an organism should be less than its wet mass) or *post hoc* (e.g., the ratio of dry to wet mass should be similar among replicates). Data validation is important for ensuring data quality and integrity by evaluating the data against a set of expectations to confirm the structure and content of the data are appropriate. In the case of data rescue, unlike most recently or currently collected data, data validation may come with the extra challenge that the original data creator or collector may be unreachable or deceased. As such, having original members of the data team (Fig. 1, Step 2; see *Team creation*) is particularly beneficial for effective data validation. Common data validation techniques include plotting the data in various ways to assist with identifying incorrect or improbable values, checking that the contents (e.g., number of unique values in a column) 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 characterize the data. In addition, many tools exist to help with the data validation process, including open-source, “point-and-click” software (e.g., OpenRefine) as well as a number of programming tools (e.g., the *assertr* and *validate* packages in R; Fischetti, 2020; van der Loo & de Jonge, 2021).

Although the exact implementation of data cleaning and validation steps will vary by dataset, many of the principles described in the *Data transfer and compilation* section are also relevant here. Validation should be conducted in as reproducible a way as possible (e.g., in a script file that can be run on the original or cleaned data files), and any errors identified should be corrected without manipulating the original (raw) data files. Importantly, 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.

6. Data archiving

Archiving data in non-proprietary formats is imperative for longevity and future accessibility. Non-proprietary software or file formats are those which do not have a copyright or trademark, and are, therefore, part of the public domain. Using non-proprietary formats ensures that anyone can access the data without needing specific (and often expensive) 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 (.xls or .xlsx).

There is a growing movement to archive data on public (and open) data repositories rather than, or in addition to, private or institutional systems (e.g., a lab hard drive). Many governments and funding agencies have recently implemented new data management protocols that either 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). With each year that passes after a publication, data that have not been publicly archived are 17% less likely to be recoverable (Vines et al., 2014; see also Tedersoo et al., 2021). 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. Once the data and metadata are compiled and validated, they should be placed in a data repository to maintain the data in a secure and retrievable format for the future. 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 (see below). 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).

There are many data repositories from which to choose (see r3data.org for a comprehensive list), with some being very generalized (e.g., Dryad, Dataverse, Figshare, Zenodo) and others catering to specific types of data (e.g., DataONE for environmental data, GenBank for genetic sequences). Data repositories tend to use a distributed (i.e., decentralized) approach to storing data and have contingency plans in place to ensure the longevity of archived datasets. Which repository to choose will also be influenced by whether the data will remain private or be made openly and publicly accessible upon upload, or sometime in the near future (Roche et al., 2014). Some repositories allow for the long-term storage of datasets regardless of whether they are made openly available (e.g., Dataverse); others require that the data be open access if they are to be hosted by the repository (e.g., Dryad). Many archives also offer an option to place an embargo, or delay, on the publication of data. Most data repositories will establish a Digital Object Identifier (DOI), a unique identifier which will remain constant for the lifetime of the object, even if the object or metadata change. If the data will be openly available, we suggest explicitly stating the terms of use, such as noting that 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 to ensure that the data meet open science standards and that their use can be tracked. Open science principles include transparency, participation and accessibility (Bartling & Friesike, 2014). 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, summarize 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 (Wilkinson et al., 2016). 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, improving the reproducibility of analyses. A comprehensive metadata file also allows 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 licenses, 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 (Carroll et al., 2020). They aim to address and encourage 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; Carroll et al., 2020). These principles are meant to begin addressing 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, and we recommend that all researchers who are rescuing datasets take these principles into consideration.

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 archive data and make them publicly accessible where appropriate. It also requires a culture change that sees best practices in data managing, archiving, and publicly sharing data become the default in publicly funded research. To achieve this goal, data sharing and accessibility need to be prioritized as critical components of the scientific enterprise. We believe that the solution to shifting the culture around data sharing is two-fold. First, there must be continued, long-term investment in data management (Mons, 2020). Such investment includes not only infrastructure but also training and support for students and personnel (Renaut et al., 2018; Soeharjono & Roche, 2021). Additionally, publishers, employers, and funding agencies must require some level of accountability from researchers to preserve data in accessible formats and, if appropriate, make the data openly available to anyone interested (Mons, 2020). Until these institutional-level paradigm shifts occur, however, smaller-scale and innovative data rescue is an integral part of environmental data curation.

Currently, training in data management and shifting regulations regarding data availability have, rightfully, focused on present and future data and data practices. With such a strong eye to the future, however, much of the 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 practices 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 recognize 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 organizations for their thoughtful discussions on data rescue, best practices in data management, and fostering more open and transparent science. We are grateful to our project partners and the data rescue interns, particularly those whose work we reference: Jenna Loesberg and Amelia Hesketh, who worked in partnership with Drs. Ellen Macdonald and Justine Karst, and Andrea Brown, who worked with Dr. Harold Eyster.

Funding for the LDP is provided through 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 funded in part 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).

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 figure. All authors revised the manuscript for publication. The authors declare no competing interests.

Figure 1. Steps in the data rescue assembly line.

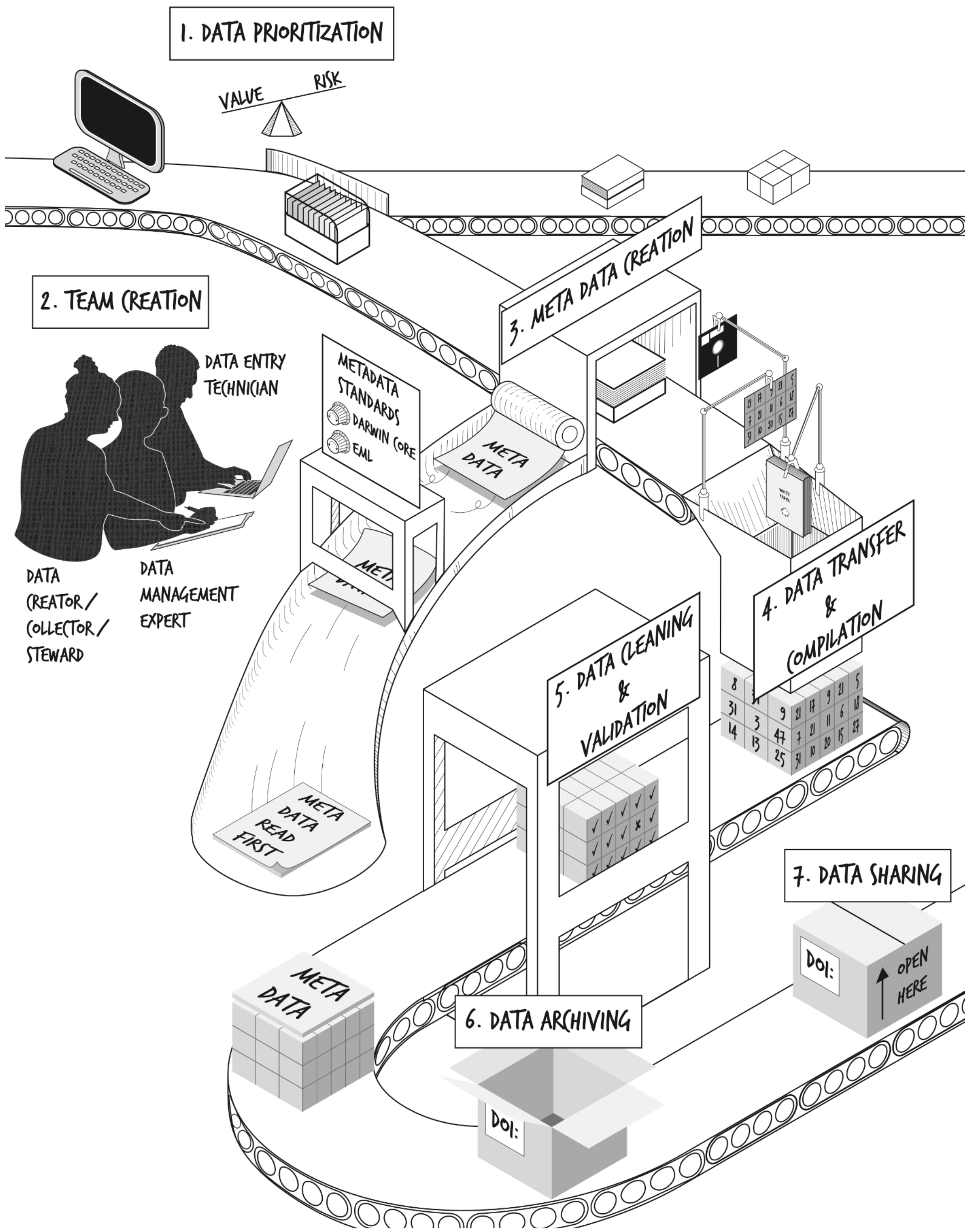


Figure 1. Steps in the data rescue assembly line. First, data must be prioritized for rescue (Step 1). After team creation (Step 2) and metadata creation (Step 3), the data must be transferred and compiled into a logical format (Step 4). After data cleaning and validation (Step 5) is complete, the finalized 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).

Figure 2. Prioritizing data for rescue: balancing the value of the data and its risk of loss.

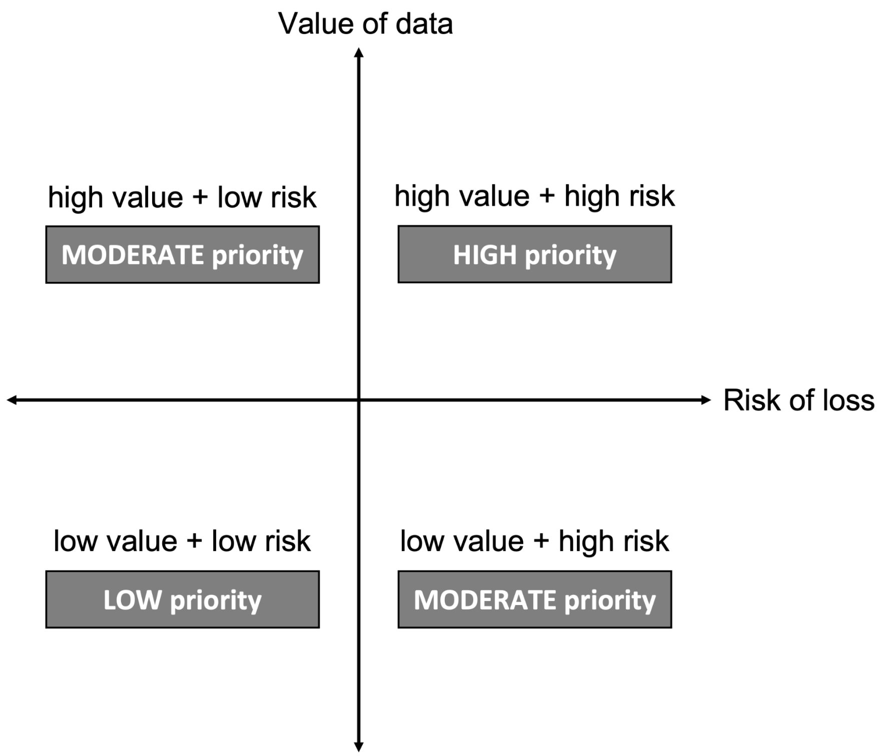


Figure 2. Prioritizing data for rescue: balancing the value of the data and its risk of loss. With many datasets in need of preservation and limited resources, the first step in the data rescue process requires developing a list of priorities for consideration and identifying relevant datasets (Fig. 1). We consider data prioritization 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 prioritization* under *Guidelines*).

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 (Jones et al., 2018). 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 (EVOSTC, 2012). Most funding went to state and federal agencies, though a few projects were awarded to universities, professional societies, consultants, and other private entities (EVOSTC, 2018).

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 (Jones et al., 2018). 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% (Jones et al., 2018). 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 nonexistent 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.

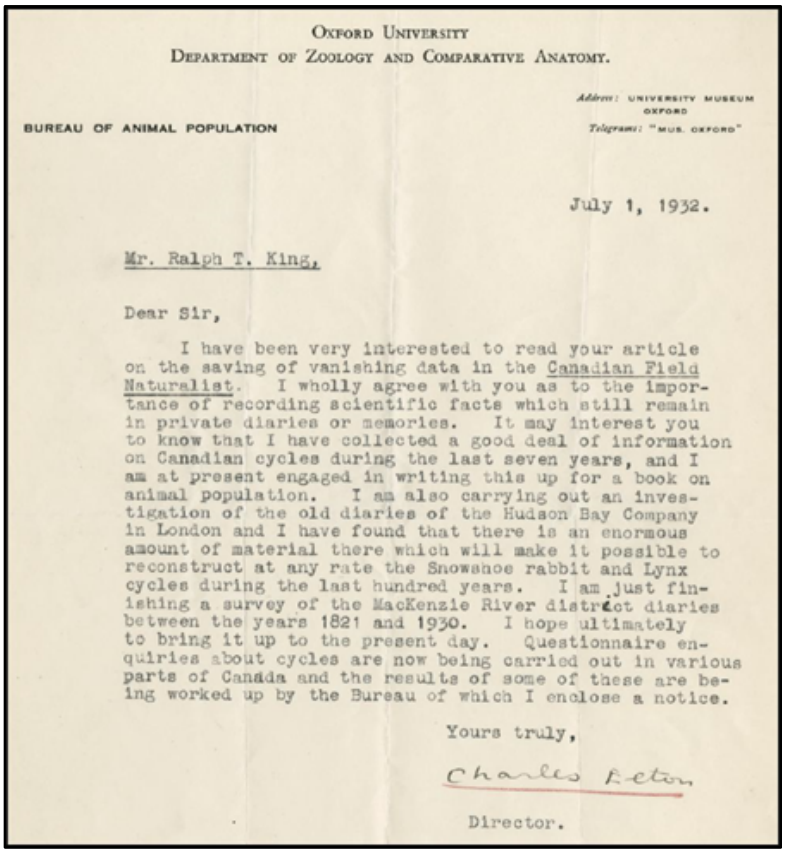
The NCEAS group also noted the reasons for their inability to recover the data. Instances in which data collectors specifically stated that the data were lost or unrecoverable were rare

(Jones et al., 2018). Instead, over 80% of reasons for unrecovered data due to a lack or failure of communication (~50% categorized as “communication lost”); the authors of the final report, however, interpret this lack of communication as an unwillingness or inability by the data owners to share data (Jones et al., 2018), highlighting the importance of proper documentation and public archiving of data for longevity.

Box 2. From fur trappers to fundamental ecological theory

Charles Sutherland Elton (1900-1991) was a British ecologist, whose major contributions include work on population cycling and community dynamics. In 1932, Elton established and became the first director of the Bureau of Animal Population at Oxford University and the inaugural editor of the *Journal of Animal Ecology* later the same year. As part of their work on population cycling, Elton and his colleague Mary Nicholson endeavoured to recover historical records on the number of Canada lynx (*Lynx canadensis*) and snowshoe hare (*Lepus americanus*) furs collected by Hudson's Bay Company (HBC) trappers in Canada (Box 2.S1; Elton & Nicholson, 1942). In an effort spanning more than 15 years, Elton and Nicholson used these and other records to collate information on trapping activities across the whole of Canada from 1886 to 1940 (and, for some areas, as far back as 1821). Much of this work was akin to data rescue, including correspondence between Elton and the original data owners (see Elton's description of the process in a letter to Ralph King in Box. 2.1; Elton & Nicholson, 1942), collation of data from different sources, as well as data cleaning and validation. This work was not only central to compiling among the longest time series of animal populations and revealing the now classical example of ~12-year population cycles in snowshoe hare and lynx abundances (Box 2.2), but has spurred an entire field of ecology (population/community cycling) and many decades of ecological research in the Canadian Arctic. This is just one, elegant example of the immense value of historical data, even those from unconventional places (like the legers of a colonial fur-trading company), and the importance of working to identify and preserve them.

Box 2.1. *Letter from Charles S. Elton to Ralph T. King, dated 01 July 1932*. In this letter, Elton expresses his interest in King's recent article on "saving vanishing data" (King, 1932), which regarded many aspects of what we are calling "data rescue" and was itself based on a paper of the same name written some three decades earlier (Haddon, 1903). This letter was provided to us courtesy of Dr. Adam T. Ford and is available through the Elton Archive at Oxford (Elton, 1932). A transcription of the letter's text is available in the Supporting Information (Box 2.S2).



OXFORD UNIVERSITY
DEPARTMENT OF ZOOLOGY AND COMPARATIVE ANATOMY.

BUREAU OF ANIMAL POPULATION

Address: UNIVERSITY MUSEUM OXFORD
Telegrams: "MUS. OXFORD"

July 1, 1932.

Mr. Ralph T. King,

Dear Sir,

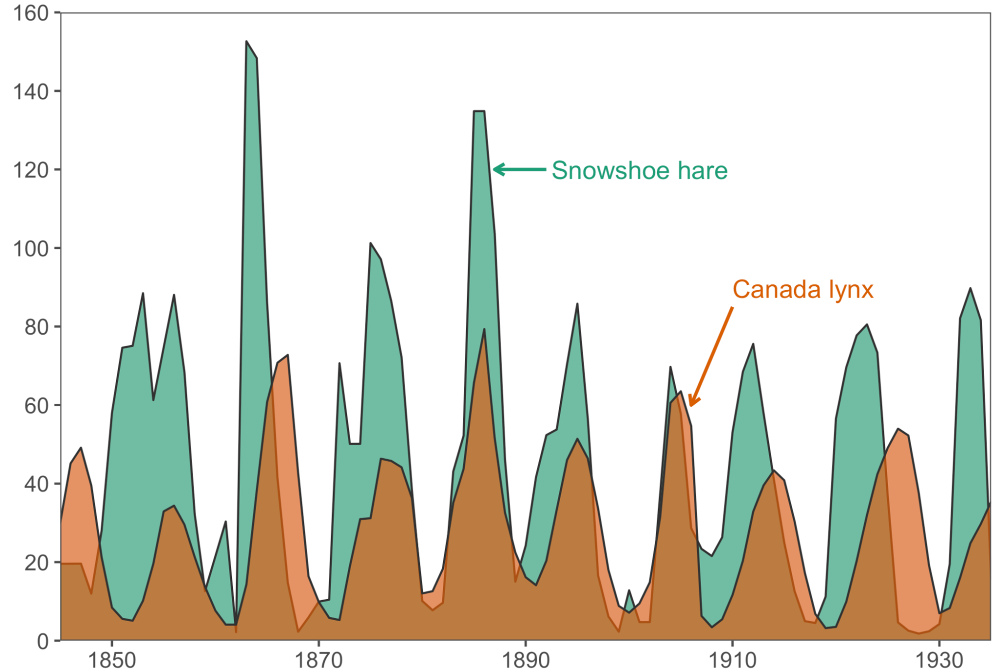
I have been very interested to read your article on the saving of vanishing data in the Canadian Field Naturalist. I wholly agree with you as to the importance of recording scientific facts which still remain in private diaries or memories. It may interest you to know that I have collected a good deal of information on Canadian cycles during the last seven years, and I am at present engaged in writing this up for a book on animal population. I am also carrying out an investigation of the old diaries of the Hudson Bay Company in London and I have found that there is an enormous amount of material there which will make it possible to reconstruct at any rate the Snowshoe rabbit and Lynx cycles during the last hundred years. I am just finishing a survey of the MacKenzie River district diaries between the years 1821 and 1930. I hope ultimately to bring it up to the present day. Questionnaire enquiries about cycles are now being carried out in various parts of Canada and the results of some of these are being worked up by the Bureau of which I enclose a notice.

Yours truly,

Charles Elton

Director.

Box 2.2. Time series of the numbers (in thousands) of Canada lynx and snowshoe hare pelts provided to the Hudson's Bay Company.



Source: recreated using data from D.R. Hundley (after MacLulich 1937)

Box 3. Recent data rescue examples from the Living Data Project.

As part of its core mission to contribute to and preserve ecological knowledge, the Living Data Project (LDP) aims to rescue valuable environmental data at risk of loss. To achieve this objective, the LDP provides training opportunities for graduate students at Canadian universities, including courses on topics and skills related to data rescue (data management, reproducibility, and collaboration) and opportunities to put these skills into practice through paid, short-term internships. The LDP partners with a variety of external organizations, including government agencies, universities, and non-profits. These partners propose potential data rescue projects, which are prioritized by a selection committee and matched to graduate student interns with the relevant skills specific to each project (e.g., with considerations for coding, database design, geospatial software, and language skills). Interns work as part of a team comprised of representatives from the partner organization, as well as postdoctoral and faculty mentors from the LDP. Below we describe two recent data rescue projects completed by LDP interns.

Seeing the Forest Data for the Trees

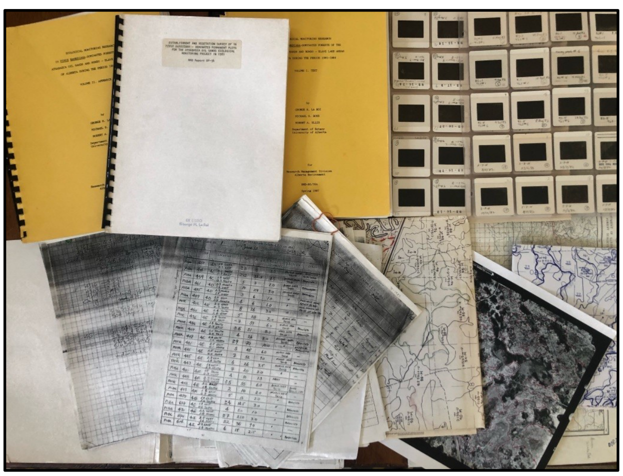
As researchers retire, they often think about the legacies they leave behind. Frequently, however, curating the data they have collected in order to cement their legacies is not at the forefront of their minds. Upon the retirement or death of a professor, students or colleagues sometimes must take the reins and piece together documents and data from decades-old research projects to ensure the data's own legacy.

Dr. George H. La Roi was a professor of forest ecology at the University of Alberta for 35 years. In a 2016 email to colleagues, he implored for assistance archiving his extensive long-term survey data from the boreal forests of northern Alberta. Before this could be accomplished, however, Dr. La Roi passed away in 2018. Upon his passing, La Roi's children bequeathed his legacy of highly valuable data to his former colleague, Dr. Ellen Macdonald, who had earlier taken over sampling some of his long-term plots. With no living data creator and much of the data in unorganized boxes containing unsorted datasheets, various documents, CD-ROMs, and picture slides, the data was at high risk of being lost. Macdonald determined she would be unable to tackle the boxes of materials alone and joined forces with her colleague, Dr. Justine Karst, who had also come into possession of some of La Roi's boxes of data by way of University's Botanic Garden. Together, they applied for an LDP data rescue internship. The value and precarious circumstances of the dataset made it a high priority for rescue.

Over the course of two data rescue internships, Jenna Loesberg and Amelia Hesketh, along with several undergraduate data entry technicians, sorted, entered, and digitized the data. They recovered data recorded at two different locations (Hondo-Slave Lake and Athabasca Oil Sands regions), both of which included measures of vascular plant cover, bryoid cover, and forest mensuration, among other datasets. Some data were stored as printed scans of hand-filled datasheets, and thus required digitization. Other data, which had already been entered and digitized, were stored in hundreds of text files which required extensive reformatting and cleaning before they could be compiled into usable datasets. Metadata also needed to be written and consolidated into one document for future reuse; while most of the data had clear documentation, some data were lost, since no documentation about the meaning of some variable names or values in a column was recovered. With this work completed, the data and metadata of this rich and expansive dataset will be archived and made publicly available through University of Alberta's Dataverse repository and eventually accompanied by a data paper.

Box 3.1. *Photograph of researchers collecting data in the Athabasca Oil Sands region of northern Alberta in 1982 at one of 16 sites established by Dr. George La Roi.* Image credit: unknown.

Box 3.2. *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. George La Roi after his passing in 2018.* Image credit: A. Hesketh.



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

The archived 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, such as word processors and tools for statistical analysis, may contain troves of raw and summary data that remain un-digitized. As a result, the reuse of any raw or summarized data from the thesis would first require data extraction and digitization.

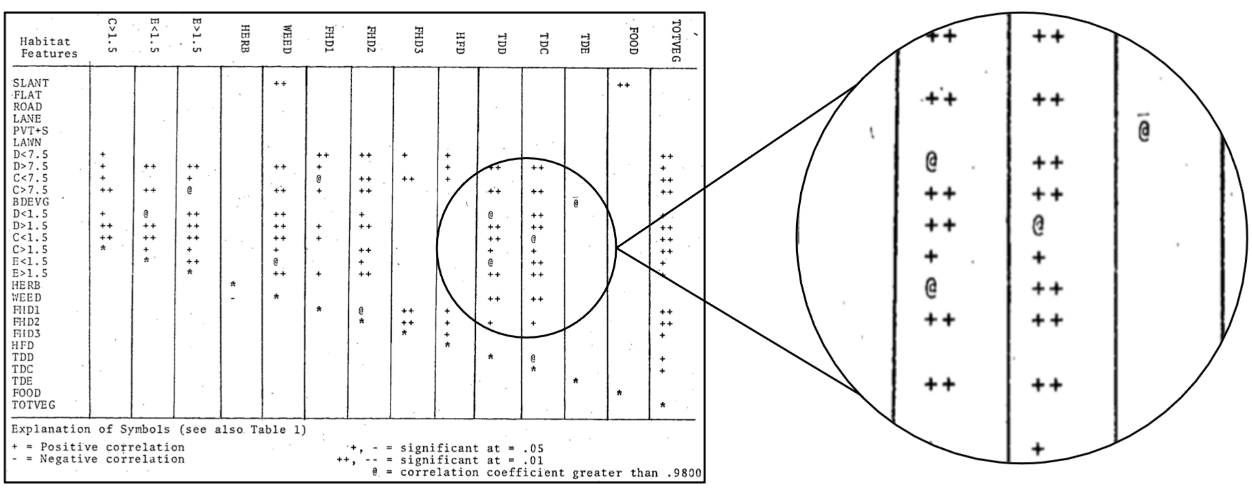
Determining how biodiversity has changed in response to human activity and land use is central to understanding the impacts of these environmental changes and predicting the potential for future declines. In a data rescue project proposed to us by a then-doctoral student, Dr. Harold Eyster, Andrea Brown worked to secure the data contained in three University of British Columbia graduate theses (Weber, 1972; Lancaster, 1976; Melles, 2000). While the specific questions and research topics differed between these theses, all three surveyed bird abundances in the same (or nearby) sites in Greater Vancouver, British Columbia, over the span of several decades, and in combination present an opportunity to establish a baseline against which to compare current and future trends (Box. 3.3). This project was prioritized by the LDP because the data were both at-risk (much of the data existed only in non-digital formats and none of the datasets are in active use) and high value (the data provide a valuable frame of reference for studying changes in urban bird diversity).

During the internship, Brown first transcribed the data from the earlier two of the theses, Weber (1972) and Lancaster (1976), which were archived as scans of typewritten documents. Among other challenges, digitization required the conversion of non-standard data types (Box 3.4) into “tidy” forms that could be used and interpreted programmatically. Data from the third thesis, Melles (2000), were made available by the original author in a Microsoft Excel® spreadsheet and so only required cleaning, manipulation, and conversion to a non-proprietary format. Later work included efforts to rationalize the datasets so that they might be used in combination with each other (e.g., standardizing column names and combining similar tables into a single file). 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. The data have been archived on the UBC Dataverse repository (Brown et al., 2021a, 2021b, 2021c) and linked with the original theses.

Box 3.3*. Comparison of the historical and current appearance of one of the sampling locations for urban bird surveys conducted in Vancouver, British Columbia, Canada*. Photographs show the view looking west from the intersection of 24th Avenue West at Wallace Street (49.251ºN, 123.191ºW).

April 1970 (Image credit: W.C. Weber) October 2021 (Image credit: C.N. Nemeth)

Box 3.4*. Example of non-standard (untidy) data to be rationalized and digitized*. This example table contains symbolic data representing the significance of correlations between habitat features. These symbols were converted to numeric factors during digitization. Reproduced with modification from Lancaster (1976; 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

Bartling S, Friesike S. 2014 *Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing*. Springer Open. (ISBN:978-2-319-00025-1)

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)

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)

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

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)

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)

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)

Buck S. 2021 Beware performative reproducibility. *Nature* **595**, 151. (doi:10.1038/d41586-021-01824-z)

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)

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)

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

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

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)

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)

Elton CS. 1932 Letter to Ralph T. King, 01 July. MS. Eng. c3328 A72, Elton Archives, Weston Library, University of Oxford.

Elton CS, Nicholson M. 1942 The ten-year cycle in numbers of the lynx in Canada. *J. Anim. Ecol.* **11**, 215-244. (https://www.jstor.org/stable/1358)

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

EVOSTC. 2018 *Exxon Valdez* Oil Spill Final and Annual Reports. *Exxon Valdez Oil Spill Trustee Council*. (https://evostc.state.ak.us/media/4291/finalandannualreports.pdf)

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)

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

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)

Haddon AC. 1903 The saving of vanishing data. *Pop. Sci. Mon.* **63**, 222-229. (https://en.wikisource.org/wiki/Popular_Science_Monthly/Volume_62/January_1903/The_Saving_of_Vanishing_Data)

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)

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)

King RT. 1932 The saving of vanishing data. *Can. Field Nat.* **46**, 108-111. (https://www.biodiversitylibrary.org/ia/canadianfieldnat1932otta/#page/134/mode/1up)

Lancaster RK. 1976 Bird communities in relation to the structure of urban habitats. Thesis. Department of Zoology, University of British Columbia. (doi:10.14288/1.0093863)

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)

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)

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)

Melles SJ. 2000 Effects of landscape and local habitat features on bird communities: a study of an urban gradient in Greater Vancouver. Thesis. Department of Forest Sciences, University of British Columbia. (doi:10.14288/1.0099590)

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)

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

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)

Powers SM, Hampton SE. 2019 Open science, reproducibility, and transparency in ecology. *Ecol. Appl.* **29**, e01822. (doi:10.1002/eap.1822)

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)

Roche DG, Berberi I, Dhane F, Lauzon F, Soeharjono S, Dakin R, Binning SA. 2021 The quality of open datasets shared by researchers in ecology and evolution is moderately repeatable and slow to change. *EcoEvoRxiv*. (doi:10.32942/osf.io/d63js)

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)

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)

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)

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)

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)

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)

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)

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)

Weber WC. 1972 Birds in cities: a study of populations, foraging ecology and nest-sites of urban birds. Thesis. Department of Zoology, University of British Columbia. (doi:10.14288/1.0101293)

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

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)

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)

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)

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)

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)