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.

* 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

1. Historical and long-term environmental datasets are imperative to understanding how natural systems respond to our changing world, setting baselines and establishing trajectories of change. Although immensely valuable, these data are ultimately at risk of being lost unless they are actively managed, curated, and eventually archived on data repositories.
2. 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 the future need for data rescue; these changes, however, do not apply retroactively. While the concept of rescuing data is not new, the term lacks a formal definition, is often conflated with other terms (i.e., data reuse), and lacks general recommendations.
3. Here, we outline seven key guidelines for effective rescue of historically-collected and unmanaged datasets. We discuss how to prioritize which datasets to rescue, form effective data rescue teams, prepare the data and related metadata, and ultimately archive and share the rescued data.
4. 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, historic data, long-term ecological data, long-term studies, open data, open science, reproducibility, transparency

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.

Abstract word count = 219 words

Main text word count (excluding abstract, figure captions, boxes, and references) = 4,931 words

Number of figures: 2

- Figure 1 = 80 words
- Figure 2 = 114 words

Number of boxes: 3

- Box 1 = 519 words (no images)
- Box 2 = 492 words (including captions; 2 images)
- Box 3 = 1,391 words (including captions; 4 images)

Number of references: 57

Why Rescue Data?

Data are among the primary units of research and scholarship. Not only are data used to help answer important questions, but they can also be used to 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 (i.e., 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). Moreover, while data collection is often targeted to a particular population, community, or location, the reuse (i.e., aggregation, collation, and synthesis) of data from different systems and contexts is essential to establishing broader ecological knowledge and informing conservation management. Yet, despite their high value and central role in research, 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 huge 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 transparency and 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 term "open science" itself was coined more than 20 years ago. However, the last decade 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; Mills et al., 2015; 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 to individuals, the scientific community and the general public are apparent (Powers & Hampton; 2019; Soeharjono & Roche, 2021). And yet, large amounts of data remain private, unavailable for reuse by other scientists, and inaccessible to researchers and the public who ultimately provided the funding and infrastructure for the data's collection. 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* below; 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 relatively little to address the issue 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 should be seen as an ethical imperative.

Here, we present guidelines for implementing data rescue; although we focus on environmental data, our guidelines are applicable more broadly. These guidelines are proposed 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). Our hope is that 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 easiest part of the process; how to implement a successful data rescue mission, however, requires a more strategic approach (Fig. 1). Some of the steps involved in data rescue are closely aligned with recommended practices in research data management (see *Metadata*, *Data Compilation*, *Validation*, *Archiving* and *Sharing* sections below). Several resources have already outlined "best" practices for data collection (Broman & Woo, 2018), management (e.g., British Ecological Society's "Data Management" Guide, 2018), and archiving (Cook et al., 2001; 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 a 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 of being lost 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, of course, 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; 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 (e.g., datasets pertaining to species at risk have higher conservation value). Additional considerations are the rarity of the data (e.g., data from an under-sampled 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 this is possible (e.g., costs of re-running major experiments or extensive surveys). Finally, a key test for the importance of a dataset is how it might be re-used in the future, with the most important datasets having many immediate potential usescenarios. 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 unreachable). Ultimately, successfully balancing the value of the data with the risk of its 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, which may include data creators, data collectors, and data stewards: *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 that act as custodians of data collected by past employees). In ecology and evolution, 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, if not more, as part of the data rescue team is imperative for a successful data rescue mission.

A *data management expert* is another key role in the data rescue process. 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 different datasets, clean and manage data, and compile previously unwritten information in detailed metadata files. Additionally, if there are any data that 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). The metadata generally describe the data collection process (including the types of data collected, methodology, and contributors, among other information), a description of all the variables in the dataset (e.g., column headings for tabular data), 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 will often inform the rest of the data rescue process and, ultimately, the structure of the compiled data.

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 match the 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; see below; Fig. 1).

The metadata file format varies (often dependent on the type of data or chosen repository). Metadata are often found in a “README” style text file. Another useful format is a text file written in Extensible Markup Language (XML; some examples and basics of XML can be found at https://www.xmlfiles.com/xml/). Tools like XML have been developed for the express purpose of writing and storing metadata and other information in a format that is both human *and* machine readable, which not only ensures that prospective end users understand the data structure and how it was created but also facilitates 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 "tags," and its description is stored between tags. In ecology, there is a set of suggested tags that should be used in such files, forming a variation of XML called Ecological Metadata Language (EML; Fegraus et al., 2005; Jones et al., 2019; see https://eml.ecoinformatics.org/).

4. Data transfer and compilation

For the data rescue team to work most 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; if there are physical copies, they 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. At a minimum, all major decisions should be documented in the metadata. For example, any edits made to the data should be done in a file separate from the original; a digital file with the untouched original data should always remain.

In structuring the data, we generally suggest following Wickham's (2014) "tidy data" principles, which consist of 3 main 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., a primary key in a relational database design; 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 other formats may be more relevant or efficient (e.g., a species by site matrix).

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 refers specifically to the process of identifying and fixing issues in the dataset, such as data entry errors or incomplete records. The importance of thorough and accurate data cleaning should not be overlooked, since, as the adage "*garbage in, garbage out*" suggests, the inference drawn from an analysis is only as strong as the inputs. In addition to common steps like correcting typographical or data 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 environmental 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 check and 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 projects, 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 as many 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 are in line with expectations following data manipulation, cross-checking data from different columns or tables for mutual compatibility (i.e., to ensure that combinations of data are within the realm of possibility), 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 depending on the nature of the dataset, many of the same general 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 during the validation process should be corrected without manipulating the original (raw) data files. Importantly, any changes made based on these checks should be well documented (e.g., as comments in the script or as notes in the metadata), as should the rationale behind the corrections. More generally, the metadata are a critical source of information for understanding the provenance of the data—that is, documentation of where the data came from, how they were collected or generated, and the steps taken to clean and compile the final dataset. Hence, thorough documentation of data validation steps is a key component of *Metadata creation* and open and reproducible data sharing.

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 a specific (and often expensive) software program 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 in proprietary formats such as Microsoft Excel files (.xls or .xlsx).

There is a strong and 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). Indeed, 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). The benefits of public archiving are clear. 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 a result, we also consider public archiving to be an essential part of data rescue, since private archiving does not mitigate the possibility that the 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). In general however, we suggest that most environmental data should be openly accessible upon archiving; exceptions include, for example, data pertaining to threatened species and considerations of Indigenous data sovereignty.

There are now a number of excellent data repositories from which to choose, with some being very generalized (e.g., Dryad, Dataverse, Figshare, Zenodo) while others cater to specific types of data (e.g., DataONE for environmental data, GenBank for genetic sequences, Global Biodiversity Information Facility (GBIF) for biodiversity data). Data repositories tend to use a distributed (i.e., decentralized) approach to storing the data and have contingency plans in place to ensure the longevity of the archived datasets (see r3data.org for a comprehensive list). Which repository to choose will also be influenced by whether the data will remain private or be made openly—and publicly—available upon upload or sometime in the near future (i.e., following an embargo period; Roche et al., 2014). Some repositories allow for the long-term preservation 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 recommend explicitly stating the terms of use for said data, 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

A final step to the data rescue workflow is to ensure that the data meet the open science standards and that their use can be tracked. Open science principles and values entail transparency, participation, and accessibility (Bartling & Friesike, 2014). These values can be addressed in different ways and, because of that, ensuring a dataset meets these standards can be overwhelming for researchers who are not trained in data management. These values can be met with a combination of actions, some of which are summarized in the FAIR and CARE principles; the first focuses on how data can be made useful and the second on how we can promote justice through responsibly sharing open data.

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 may also help with the automatic discoverability of datasets. As mentioned in the *Data archiving* section, tagging a dataset with a DOI makes it trackable and citable, which improves the reproducibility of analyses. Many online data repositories provide DOIs for datasets, and they are crucial to connect the actual dataset to its metadata (which will also be registered under the same DOI). 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.

Ongoing data rescue initiatives

Data rescue is not a new concept (e.g., Hawkins et al., 2013; Specht et al., 2018), and a number of examples have been noted in both the scientific and grey literature (e.g., Box 2; Norton et al., 2000; Hawkings et al., 2013; Kelly et al., 2016; Specht et al., 2018; Knockaert et al., 2019). That said, the approach lacks a formal definition and can be conflated with other terms, such as data reuse. It also lacks general guidelines and best practices, of which we have offered a brief overview in this paper.

Some data rescue efforts have embraced community science, using crowdsourcing platforms such as Zooniverse to facilitate data rescue (e.g., Unearthing Michigan Ecological Data). Additionally, there are currently a few organizations focused on preserving ecological data at risk. For example, the Canadian Institute for Ecology and Evolution (CIEE/ICEE) started the Living Data Project (LDP) in 2018 with a mission to rescue and breathe new life into languishing ecological, evolutionary, and environmental datasets (Box 3). Other organizations practicing organized data rescue include (but are not limited to) the Atmospheric Circulation Reconstructions over the Earth (ACRE) and the International Environmental Data Rescue Organization (IEDRO).

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. While there has been movement in this direction, we are still far from the ideal. To achieve this goal, data sharing and accessibility need to be prioritized as a critical component 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; Ritchie, 2021). Such investment includes not only infrastructure but also training and support for students and personnel (Soeharjono & Roche, 2021). Additionally, publishers, employers and funding agencies must require some level of accountability from researchers to preserve data in accessible, non-proprietary formats and, if appropriate, make those data openly available to anyone interested (Mons, 2020). Until these large, 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 past data while also providing additional, less tangible benefits. In the LDP, 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 multiple 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 would like to thank LDP members and partner organizations for their thoughtful discussions on data rescue, best practices in data management, and fostering a more open and transparent science. We are grateful to all of our project partners and the data rescue interns, and in particular to those whose work we reference in this manuscript: Jenna Loesberg and Amelia Hesketh, who worked in partnership with Drs. Ellen Macdonald and Justine Karst at the

University of Alberta, and Andrea Brown, who worked with Dr. Harold Eyster at the University of British Columbia.

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).

Figure 1. Steps in the data rescue process.

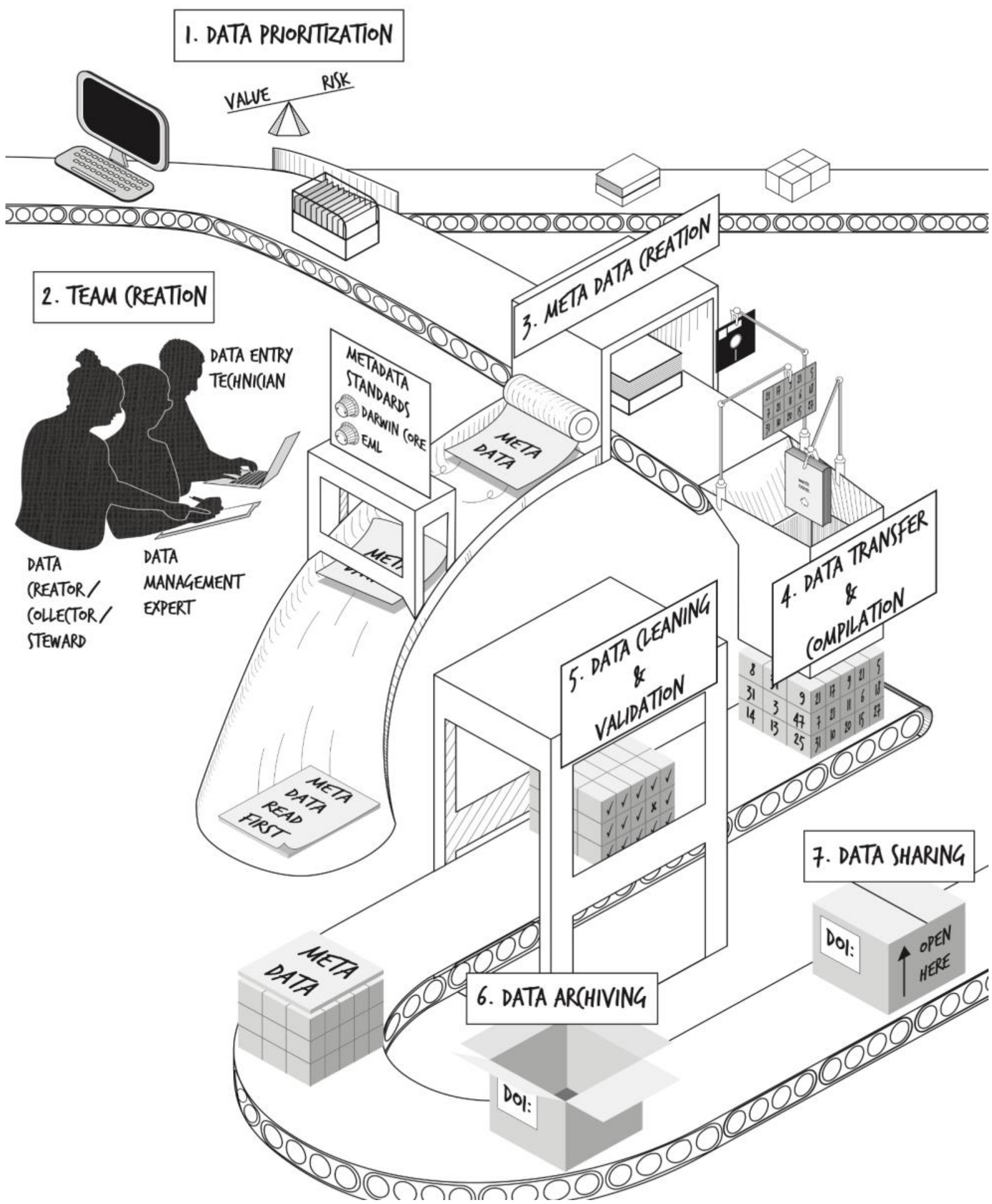


Figure 1. Steps in the data rescue process. First, data must be prioritized for rescue (Step 1). After team creation (Step 2) and metadata creation (Step 3), the data must be transferred and compiled into a consistent and effective format (Step 4). After data cleaning and validation (Step 5) is complete,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

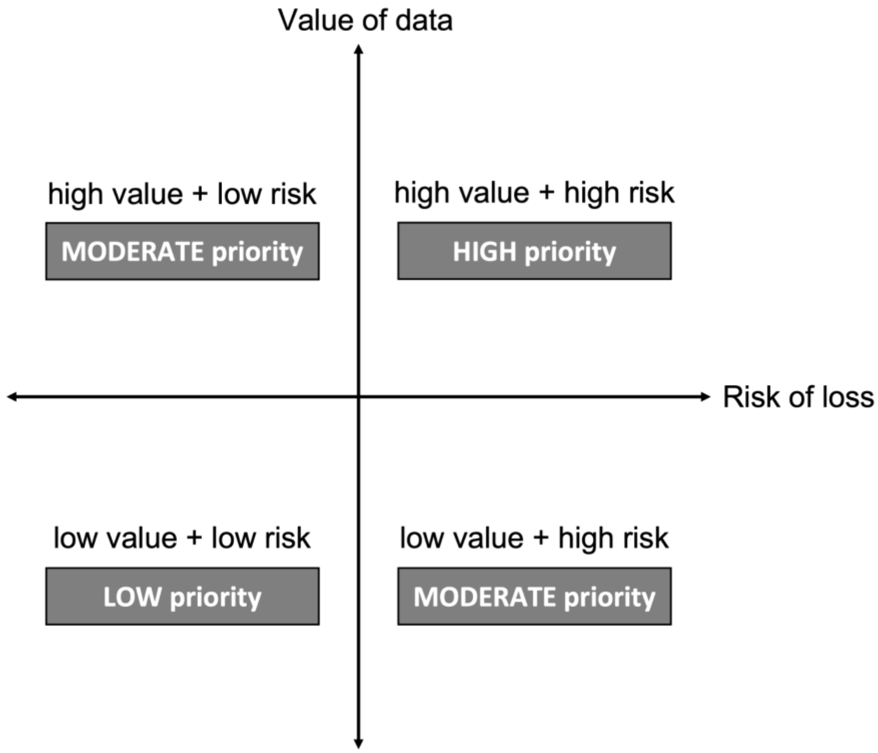


Figure 2. Prioritizing data for rescue: balancing the value of the data and its risk of loss. With many datasets in need of preservation and limited resources, the first step in the data rescue process requires developing a list of priorities for consideration and identifying relevant datasets (Fig. 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*). While the process of prioritization is inherently subjective, we suggest that considerations of value and risk can provide a useful heuristic for practitioners looking to best target their time and effort.

Box 1. Spilt oil, spent money, and lost data: Exxon-Valdez oil spill as a case study on the costs of data loss.

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 the 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 the amount of funds spent on "Research, Monitoring, and General Restoration" during 1992-2010 fiscal years was $151.2 million USD (EVOSTC, 2012). The majority of funding went to state and federal agencies, though a few projects were awarded to universities, professional societies, consultants, and other private entities (EVOSTC, 2018).

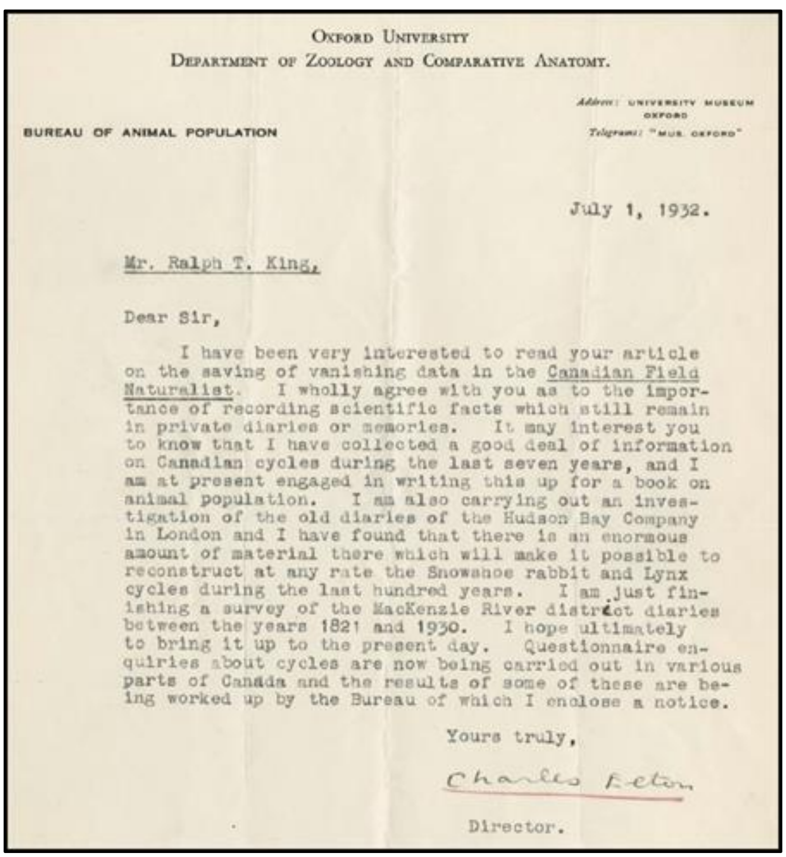
From 2012-2014, a group of researchers from the National Center for Ecological Analysis & Synthesis (NCEAS) worked to recover the historical datasets funded by EVOSTC, focusing specifically on data collected between 1989-2010 (Jones et al., 2018). Of the 419 projects determined to have been funded by 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 calculate a rough estimate of money spent on research for which the data are not recoverable (70% of datasets): approximately **$105 million USD was spent collecting data which are no longer recoverable and, therefore, effectively lost to science.** While we do not know the distribution of years from which data were recovered or how money was divided by year, $105 million USD 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. Similarly, this valuation does not include any of the nearly $50 million USD spent towards "Scientific Management, Public Information & Administration" (EVOSTC, 2012).

The group tasked with recovering these historic datasets 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 datasets which were unrecovered were lost due to a lack or failure of communication (~50% categorized as "communication lost"); the authors of the final report, however, interpret much of 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 putting datasets in publicly available data repositories for longevity.

Box 2. From fur trappers to fundamental ecological theory: how a data rescue effort shaped our understanding of population cycling.

Charles Sutherland Elton (29 May 1900 – 01 May 1991) was a British ecologist, whose major contributions included work on population cycling and community dynamics. In 1932, Elton established and became the first director of the Bureau of Animal Population (BAP) 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*) furs collected by Hudson's Bay Company (HBC) trappers in Canada (Box 2.S1; Elton & Nicholson, 1942). In an effort that spanned 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 as far back as 1821 for areas in the Mackenzie River District, Northwest Territories. 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 one of the longest time series of animal populations and revealing the now classical example of ~12-year population cycles in snowshoe hare and Canada 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 (Bureau of Animal Population, Oxford University) to Ralph T. King (SUNY College of Environmental Science and Forestry), 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). Elton goes on to describe his efforts to reconstruct time series of hares and lynx from HBC records. This letter is not only an important historical artifact, but also highlights a "tradition" of data rescue that dates to the formalization of ecology as a discipline. 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,

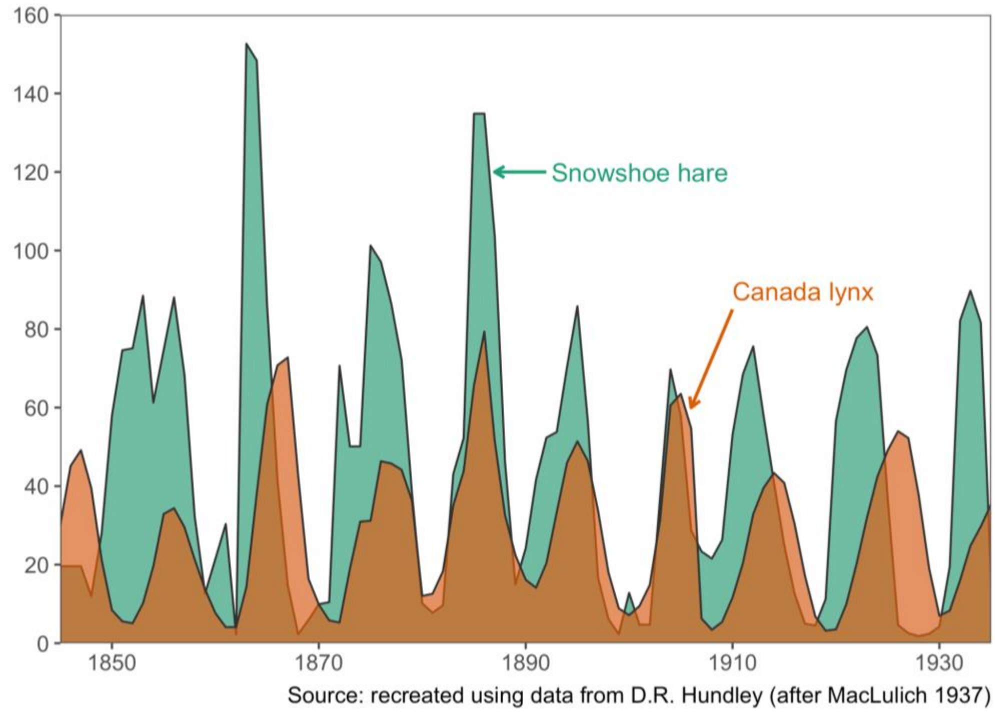
I have been very interested to read your article on the saving of vanishing data in the Canadian Field Naturalist. I wholly agree with you as to the importance of recording scientific facts which still remain in private diaries or memories. It may interest you to know that I have collected a good deal 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.



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 ecological and environmental data at risk of being lost. 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

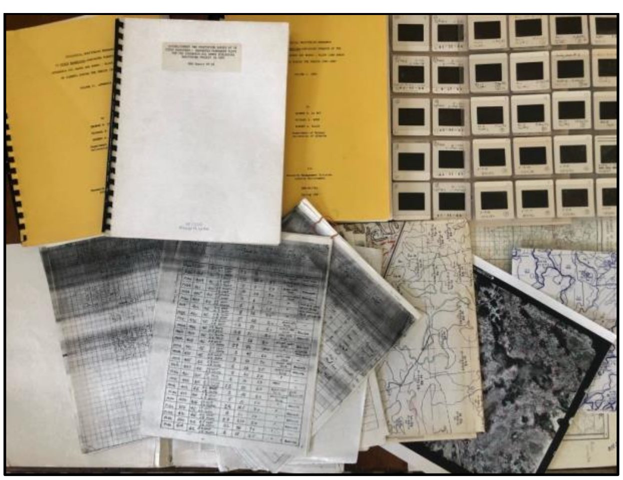
As researchers retire, they often think about the legacies they leave behind. Frequently, however, curating the data they have collected in order to cement their legacies is not at the forefront of their minds. Upon the retirement or death of a professor, students or colleagues often must take the reins 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 2016, he composed an email to colleagues asking for help archiving his extensive long-term survey data from the boreal forests in Alberta. Before this could be accomplished, however, Dr. La Roi passed away in 2018. Upon his passing, La Roi's children bequeathed much of 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 with 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 knew she would not be able to tackle the boxes of materials on her own and joined forces with another University of Alberta colleague, Dr. Justine Karst, who had also come into possession of some of La Roi's boxes of data by way of University of Alberta's Botanic Garden. Together, they wrote an application for an LDP data rescue internship. With the data being highly valuable long-term data and also at a high risk of loss, this dataset was deemed to be of high priority for rescue.

Over the course of two data rescue internships, graduate students Jenna Loesberg and Amelia Hesketh, along with a handful of undergraduate data entry technicians, sorted, entered, and digitized the data. They determined that there were data from two different locations—the Hondo-Slave Lake region and the Athabasca Oil Sands region—both of which included data on vascular plant cover, bryoid cover, and forest mensuration, among other datasets. Some data were found only on printed-out scans of hand-written datasheets and needed to be entered into a digital format. 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, as no documentation about the meaning of the variable names or the values in the column could be found or determined. With all of 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 hopefully published as a data paper.

Box 3.1. *Photograph of researchers collecting data in the Athabasca Oil Sands region of northern Alberta in 1982. This is one of 16 sites established by Dr. George La Roi in the region in the 1980s to study seasonal and annual dynamics of boreal forests.* 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, though not fully realized, 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, contain troves of raw and summary data that have not been digitized and archived, and so remain inaccessible to present-day researchers. As a result, the reuse of any raw or summarized data from the thesis would first require data extraction and digitization.

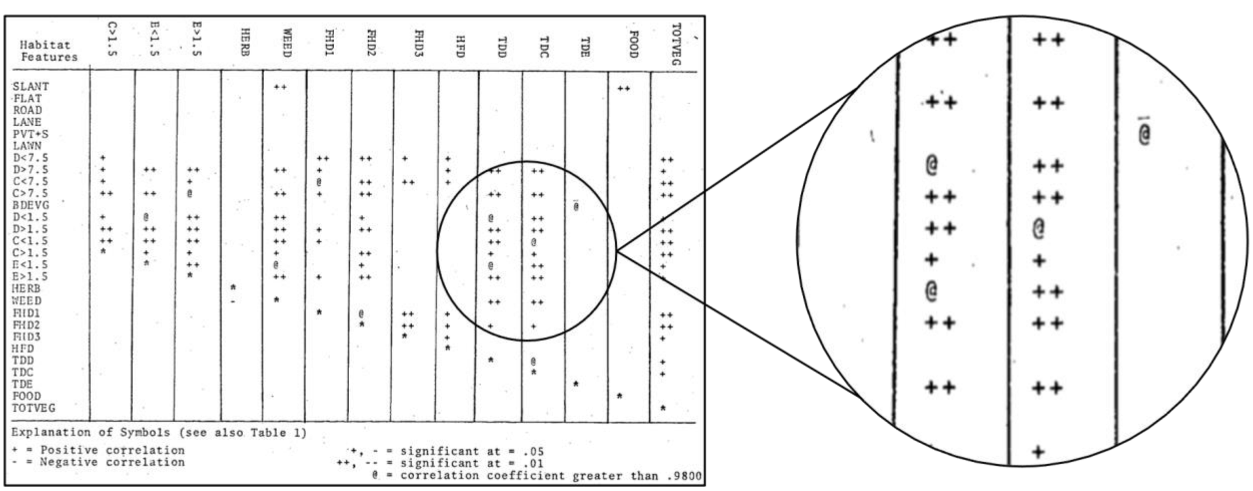
Urban areas have expanded in size, number, and human population density in recent decades, accompanied by changes in the abundance and diversity of bird populations that inhabit these regions. Determining how biodiversity has changed in response to historical changes in human activity and land use is 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, LDP intern Andrea Brown worked to secure the data contained in three University of British Columbia graduate theses (Weber, 1972; Lancaster, 1976; Melles, 2000), with a particular focus on data pertaining to surveys of bird abundances at various locations around Greater Vancouver, British Columbia, Canada. While the specific questions and research topics differed between these theses, the fact that all three surveyed the same (or nearby) sites in Greater Vancouver over the span of several decades means that, in combination, they present an opportunity to establish a baseline against which to compare current and future trends (Box. 3.3 shows an example of the change in conditions at one of the sites sampled by Weber (1972)). This project was identified as a priority for 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 her internship, Brown first worked to transcribe the data from the earlier two of the theses, Weber (1972) and Lancaster (1976), which were archived as scans of typewritten documents and did not have data available in digital form. Among other challenges, digitization required the conversion of non-standard data types (see, e.g., 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 and 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 other formatting, or combining similar or related 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 dataset 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, Eyster, & Lancaster, 2021; Brown, Eyster, & Melles, 2021; Brown, Eyster, & Weber 2021) 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). The historical reference is reproduced from Weber (1972); the contemporary image is shared with permission from the photographer.

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.

Borgman, C. L. (2018). Open data, grey data, and stewardship: universities at the privacy frontier. *Berkeley Technology Law Journal*, 33, 365-412. https://doi.org/10.15779/Z38B56D489

British Ecological Society. (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

Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. *American Statistician*, 72, 2-10. https://doi.org/10.1080/00031305.2017.1375989

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

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

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

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

Buxton, R. T., Nyboer, E. A., Pigeon, K. E., Raby, G. D., Rytwinski, T., Gallagher, A. J., Schuster, R., Lin, H.-Y., Fahrig, L., Bennett, J.R, Cooke, S.J., & Roche, D.G. (2021). Avoiding wasted research resources in conservation science. *Conservation Science and Practice*, 3(2), e329. https://doi.org/10.1111/csp2.329

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE principles for Indigenous data governance. *Data Science Journal*, 19, 43. https://doi.org/10.5334/dsj-2020-043

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

Codd, E. F. (1990). *The Relational Model for Database Management: Version 2.* Addison-Wesley Longman Publishing.

Cook, R. B., Olson, R. J., Kanciruk, P., & Hook, L. A. (2001). Best practices for preparing ecological data sets to share and archive. *Bulletin of the Ecological Society of America*, 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. https://doi.org/10.1007/BF01096442

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

Elton, C. S., & Nicholson, M. (1942). The ten-year cycle in numbers of the lynx in Canada. *Journal of Animal Ecology*, 11, 215-244. https://www.jstor.org/stable/1358

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

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

Fegraus, E. H., Andelman, S., Jones, M. B., & Schildhauer, M. (2005). Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation. *Bulletin of the Ecological Society of America*. 86(3), 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, C. N., 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. https://doi.org/10.1371/journal.pone.0118581

Haddon, A. C. (1903). The saving of vanishing data. *Popular Science Monthly*, 63, 222-229. https://en.wikisource.org/wiki/Popular_Science_Monthly/Volume_62/January_1903/The_Saving_of_Vanishing_Data

Hawkings, S. J., Firth, L. B., McHugh, M., Poloczanska, E. S., Herbert, R. J. H., Burrows, M. T., Kendall, M. A., Moore, P. J., Thompson, R. C., Jenkins, S. R., Sims, D. W., Genner, M. J., & Mieszkowska, N. (2013). Data rescue and re-use: recycling old information to inform new policy concerns. *Marine Policy*, 42, 91-98. https://doi.org/10.1016/j.marpol.2013.02.001

Jones, M. B., 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, M. B., 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. https://doi.org/10.5063/F11834T2

Kelly, G., Easterday, K., Rapucciuolo, G., Koo, M. S., McIntyre, P., & Thorne, J. (2016). Rescuing and sharing historical vegetation data for ecological analysis: the California Vegetation Type Mapping Project. *Biodiversity Informatics*, 11, 40-62. https://core.ac.uk/download/pdf/162636907.pdf

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

Knockaert, C., Tyberghein, L., Goffin, A., Vanhaecke, D., Ong'anda, H., Wakwabi, E. O., & Mees, J. (2019). Biodiversity data rescue in the framework of a long-term Kenya-Belgium cooperation in marine sciences. *Scientific Data* 6(85). https://doi.org/10.1038/s41597-019-0092-8

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

Lennox, R.J., Harcourt, R., Bennett, J.R., Davies, A., Ford, A.T., Frey, R.M., …, & Cooke, S. J. (2020). A novel framework to protect animal data in a world of biosurveillance. *BioScience*, 70, 468-476. https://doi.org/10.1093/biosci/biaa035

van der Loo, M.P.J., & de Jonge, E. (2021). Data validation infrastructure for R. *Journal of Statistical Software*, 97, 1–31. https://doi.org/10.18637/jss.v097.i10

McClenachan, L., Ferretti, F., & Baum, J. K. (2012). From archives to conservation: why historical data are needed to set baselines for marine animals and ecosystems. *Conservation Letters*, 5, 349-359. https://doi.org/10.1111/j.1755-263X.2012.00253.x

Melles, S. J. (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. https://dx.doi.org/10.14288/1.0099590

Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B. & Stafford, S. G. (1997). Nongeospatial metadata for the ecological sciences. *Ecological Applications*, 7, 330-342.

Mills, J. A., Teplitsky, C., Arroyo, B., Charmantier, A., Becker, P. H., Birkhead, T. R., Bize, P., Blumstein, D. T., Bonenfant, C., Boutin, S., Bushuev, A., Cam, E., Cockburn, A., Côté S. D., Coulson, J. C., Daunt, F., Dingemanse, N. J., Doligez, B., Drummond H., Espie, R. H. M., et al. (2015). Archiving primary data: solutions for long-term studies. *Trends in Ecology and Evolution*, 30, 581-589. https://doi.org/10.1016/j.tree.2015.07.006

Mons, B. (2020). Invest 5% of research funds in ensuring data are reusable. *Nature*, 578(7796), 491. https://doi.org/10.1038/d41586-020-00505-7

Norton, D. C., Assel, R. A., Meyers, D., Hibner, B. A., Morse, N., Trimble, P. J., Cronk, K., & Rubens, M. (2000). Great Lakes ice data rescue project (Technical memorandum GLERL-117). Great Lakes Environmental Research Laboratory, National Oceanographic and Atmospheric Administration (NOAA). https://repository.library.noaa.gov/view/noaa/11024

O'Dea, R. E., Parker, T. H., Chee, Y. E., Culina, A., Drobniak, S. M., Duncan, D. H., ... & Nakagawa, S. (2021). Towards open, reliable, & transparent ecology and evolutionary biology. *BMC Biology*, 19(1), 1-5. https://doi.org/10.1186/s12915-021-01006-3

Powers, S. M., & Hampton, S. E. (2019). Open science, reproducibility, and transparency in ecology. *Ecological Applications*, 29, e01822. https://doi.org/10.1002/eap.1822

Ritchie, H. (2021). COVID's lessons for climate, sustainability and more from Our World in Data. *Nature*, 598:9. https://doi.org/10.1038/d41586-021-02691-4

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

Roche, D. G., Kruuk, L. E. B., Lanfear, R., & Binning, S. A. (2015). Public data archiving in ecology and evolution: how well are we doing? *PLoS Biology*, 13, e1002295. https://doi.org/10.1371/journal.pbio.1002295

Roche, D. G., Lanfear, R., Binning, S. A., Haff, T. M., Schwanz, L. E., Cain, K. E., Kokko, H., Jennions, M. D., & Kruuk, L. E. (2014). Troubleshooting public data archiving: suggestions to increase participation. *PLoS Biology*, 12(1), e1001779. https://doi.org/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 Research*, 5, 781. https://doi.org/10.12688/f1000research.8422.1

Specht, A., Bolton, M. P., Kingsford, B., Specht, R. L., & Belbin, L. (2018). A story of data won, data lost and data re-found: the realities of ecological data preservation. *Biodiversity Data Journal*, 6, e29073. https://doi.org/10.3897/BDJ.6.e28073

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. https://doi.org/10.1371/journal.pone.0234172

Soeharjono, S., & Roche, D. R. (2021). Reported individual costs and benefits of sharing open data among Canadian academic faculty in ecology and evolution. *BioScience*, biab024. https://doi.org/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. *Scientific Data*, 8, 192. https://doi.org/10.1093/biosci/biab024

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., Gilbert, K. J., Moore, J., Renault, S., & Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current Biology*, 24, 94-97. https://doi.org/10.1016/j.cub.2013.11.014

Walter, M., & Suina, M. (2018). Indigenous data, indigenous methodologies and indigenous data sovereignty. *International Journal of Social Research Methodology*, 22, 233-243. https://doi.org/10.1080/13645579.2018.1531228

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

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, 59(10). http://dx.doi.org/10.18637/jss.v059.i10

Wilkinson, M. D., Dumontier, M., Aalbersberg, IJ. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., Bonino de Silva Santos, L., Bourne, P. E., Bouwman, J., Brooke, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Willis, K. J., Araùjo, M. B., Bennett, K. D., Figueroa-Rangel, B., Freud, C. A., & 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. *Philosophical Transactions of the Royal Society B*, 362, 175-187. https://doi.org/10.1098/rstb.2006.1977

White, E. P., Baldridge, E., Brym, Z. T., Locey, K. J., McGlinn, D. J. & Supp, S. R. (2013). Nine simple ways to make it easier to (re) use your data. *Ideas in Ecology and Evolution*, 6, 1–10. https://doi.org/10.4033/iee.2013.6b.6.f

Whitlock, M. C. (2011). Data archiving in ecology and evolution: best practices. Trends in Ecology and Evolution, 26, 61-65. https://doi.org/10.1016/j.tree.2010.11.006

Yenni, G. M., Christensen, E. M., Bledsoe, E. K., Supp, S. R., Diaz, R. M., White, E. P., & Ernest, S. M. (2019). Developing a modern data workflow for regularly updated data. *PLoS Biology*, 17(1), e3000125. https://doi.org/10.1371/journal.pbio.3000125