JESLIB 2013; 2(2): 3-16
doi:10.7191/jeslib.2013.1024

Journal of eScience Librarianship
putting the pieces together: theory and practice

Common Errors in Ecological Data Sharing

Karina E. Kervin,[1] William K. Michener,[2] Robert B. Cook[3]

[1] University of Michigan, Ann Arbor, MI, USA
[2] University of New Mexico, Albuquerque, NM, USA
[3] Oak Ridge National Laboratory, Oak Ridge, TN, USA

Abstract

Objectives: (1) to identify common errors in data organization and metadata completeness that would preclude a "reader" from being able to interpret and re-use the data for a new purpose; and (2) to develop a set of best practices derived from these common errors that would guide researchers in creating more usable data products that could be readily shared, interpreted, and used.

Methods: We used directed qualitative content analysis to assess and categorize data and metadata errors identified by peer reviewers of data papers published in the Ecological Society of America's (ESA) *Ecological Archives*. Descriptive statistics provided the relative frequency of the errors identified during the peer review process.

Results: There were seven overarching error categories: Collection & Organization, Assure, Description, Preserve, Discover, Integrate, and Analyze/Visualize. These categories represent errors researchers regularly make at each stage of the Data Life Cycle. Collection & Organization and Description errors were some of the most common errors, both of which occurred in over 90% of the papers.

Conclusions: Publishing data for sharing and reuse is error prone, and each stage of the Data Life Cycle presents opportunities for mistakes. The most common errors occurred when the researcher did not provide adequate metadata to enable others to interpret and potentially re-use the data. Fortunately, there are ways to minimize these mistakes through carefully recording all details about study context, data collection, QA/QC, and analytical procedures from the beginning of a research project and then including this descriptive information in the metadata.

Introduction

Data are increasingly being recognized as important products of the scientific enterprise (U.S. GAO 2007; OSTP 2013) and funding agencies such as the U.S. National Institutes of Health and U.S. National Science Foundation. Both agencies now require that proposals include plans describing how data will be shared and managed (NIH 2003, NSF 2011). Similarly, professional societies and journals have endorsed the principles of the Joint Data Archiving Policy (e.g., Moore et al. 2010, Rausher et al. 2010, Riesenberg et al. 2010, Whitlock et al. 2010, Wenburg 2011) and are encouraging authors to archive primary data sets and metadata in an appropriate public archive (e.g., Dryad, TreeBASE, GenBank, Protein Databank, etc.). In order for the data to be interpreted, shared, and re-used, it must also be accompanied by metadata that describe the scientific context

Correspondence to Karina E. Kervin: kkervin@umich.edu
Keywords: data publication, data management, data sharing, best practices

for the data as well as how the data were generated, organized, quality assured, and preserved (Michener et al. 1997).

The process of publishing data and metadata is relatively new to scientists in many domains. The Ecological Society of America's (ESA) data papers represent a unique type of article that the ESA has published since 2005. ESA's *Ecology* publishes the abstract describing the data paper and *Ecological Archives* publishes the comprehensive data sets and accompanying metadata that describe the content, context, quality, and structure of the data. *Ecological Archives* provides long-term access to data papers which authors are encouraged to periodically update to facilitate secondary data use and analysis. Data papers undergo extensive peer review to assess the submission's overall quality and significance to the ecological sciences as well as additional technical review of the data and metadata to ensure a high standard of usability.

The overall goal of this paper was to provide a detailed case study of common errors observed when researchers prepare data and documentation for sharing and archiving. The findings were derived from ecology but are applicable for other research disciplines that require data management for long-term archiving, as well as libraries, data librarians, and archivists that may play a role in supporting researchers. This study analyzed peer reviews of 53 ESA data papers published in *Ecology* and *Ecological Archives* between August 2005 and May 2012. The principal objectives of this study were: (1) to identify common errors in data organization and metadata completeness that would preclude a "reader" from being able to interpret and re-use the data for a new purpose (e.g. study repeatability, synthesis, or meta-analysis); and (2) to develop a set of best practices derived from these common errors that would guide researchers across disciplines in creating more usable data products that could be readily shared, interpreted, and used.

Methods and Procedures

Data Collection

Data papers included both the data files and associated metadata that the author(s) submitted to the *Ecological Archives*. Data paper authors were required to follow the *Ecological Archives* metadata content standard which is based on the format described in Michener et al. (1997) and includes a comprehensive list of elements that, if adequately described in the data paper, should be sufficient to allow researchers unfamiliar with the data set to effectively interpret and reuse the data.

Two or more peer reviewers who the editor of *Ecological Archives* considered subject-matter experts in the topic of the paper reviewed each submission. Peer reviewers focused on four aspects of the paper (ESA Archives 2012): "1. Importance and interest to *Ecological Archives'* users and readers. 2. Scientific and technical soundness of the database. 3. Originality. 4. Degree to which metadata fully describe the content, context, quality, and structure of the database." Reviewers were encouraged to specifically comment on: metadata presentation and completeness; data organization, quality, and integrity; methods; study design; errors; and citations. The editor for *Ecological Archives* evaluated the reviews and decided whether to accept the data paper or allow for resubmission after the author(s) addressed the reviewers' comments. Revised data papers were further evaluated by the editor and published if the revisions were deemed suitable in responding to the reviewers' comments.

Ecology and *Ecological Archives* published all 53 data papers used in this analysis after requested revisions were completed, including satisfactorily addressing all issues identified by the reviewers (Table 1). A total of 104 peer reviews of all published data papers provided the data that were analyzed for this paper (Table 1). Peer-review com-

Table 1: Number of data papers published by year, including total number of reviews per year and average number of reviews per data paper. Papers evaluated in 2012 represent a partial year.

Calendar Year	Number of Data Papers	Number of reviews	Average number of reviews
2004	1	2	2.0
2005	4	8	2.0
2006	1	2	2.0
2007	8	16	2.0
2008	9	17	1.9
2009	8	17	2.1
2010	5	10	2.0
2011	15	28	1.9
2012*	2	4	2.0
Total	53	104	2.0

ments of rejected data papers were not available for analysis; in most such cases, the editor rejected the data papers as inappropriate for *Ecological Archives*, and the data papers were not sent to peer reviewers.

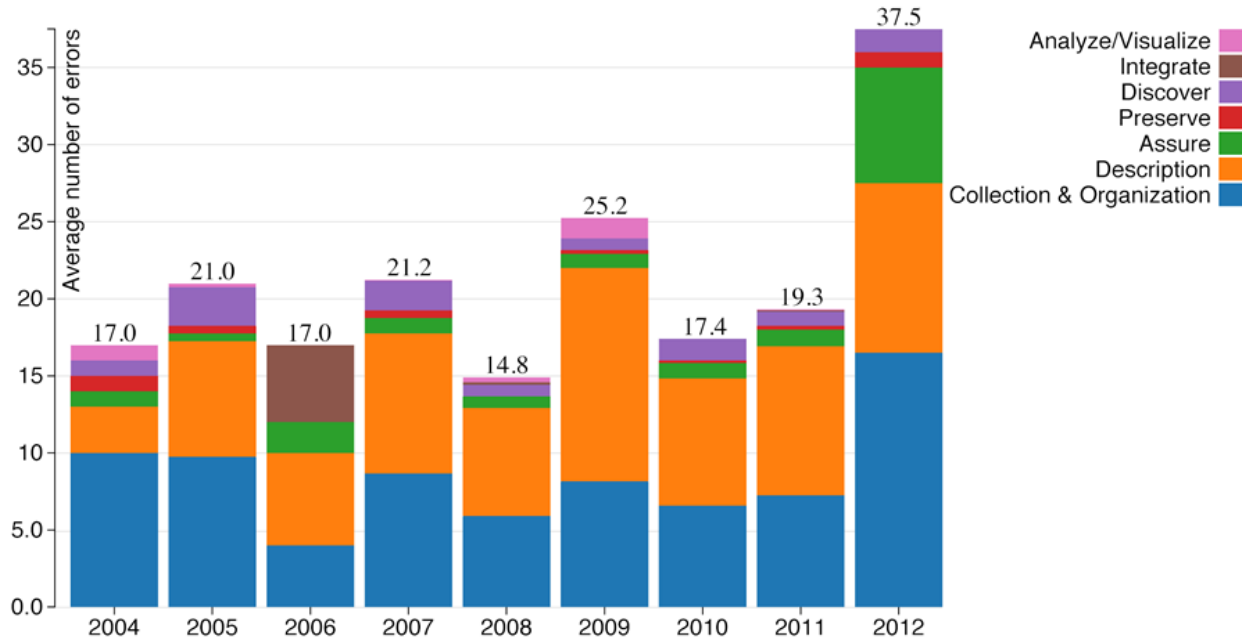
The number of data papers submitted generally increased over time (see Table 1). Researchers submitted few data papers during the first few years. The number of data papers submitted increased in the fourth year, with the peak in 2011, the last full year analyzed.

Data Analysis

Directed qualitative content analysis (Zhang & Wildemuth 2009) was used to assess data and metadata errors identified by peer reviewers of papers published in the *Ecological Archives*. Descriptive statistics provided the relative frequency of the various errors identified during the peer-review process.

Analysis of the data paper reviews consisted of qualitative coding of errors followed by quantitative analysis of those codes. First, five data papers were selected at random, and reviewer-identified errors were identified and listed. Second, those errors were grouped into the Data Life Cycle elements described by Michener and Jones (2012): (1) Collection & Organization; (2) Assure (including quality assurance and quality control); (3) Description (i.e., ascribing metadata to the data); (4) Preserve; (5) Discover; (6) Integrate; and (7) Analyze/Visualize. While the Data Life Cycle also includes an eighth element (Planning), these types of errors were not apparent in the reviewer's comments. Third, errors were assigned to more detailed categories based on the metadata elements identified by Michener and others (1997). For a complete list of error categories, see Appendix A. Finally, the reviews of the remaining 48 data papers were analyzed by categorizing the reviewer-identified errors

Figure 1: Average number of errors per paper by year by Data Life Cycle Element Category



into individual categories. When necessary, new categories were created.

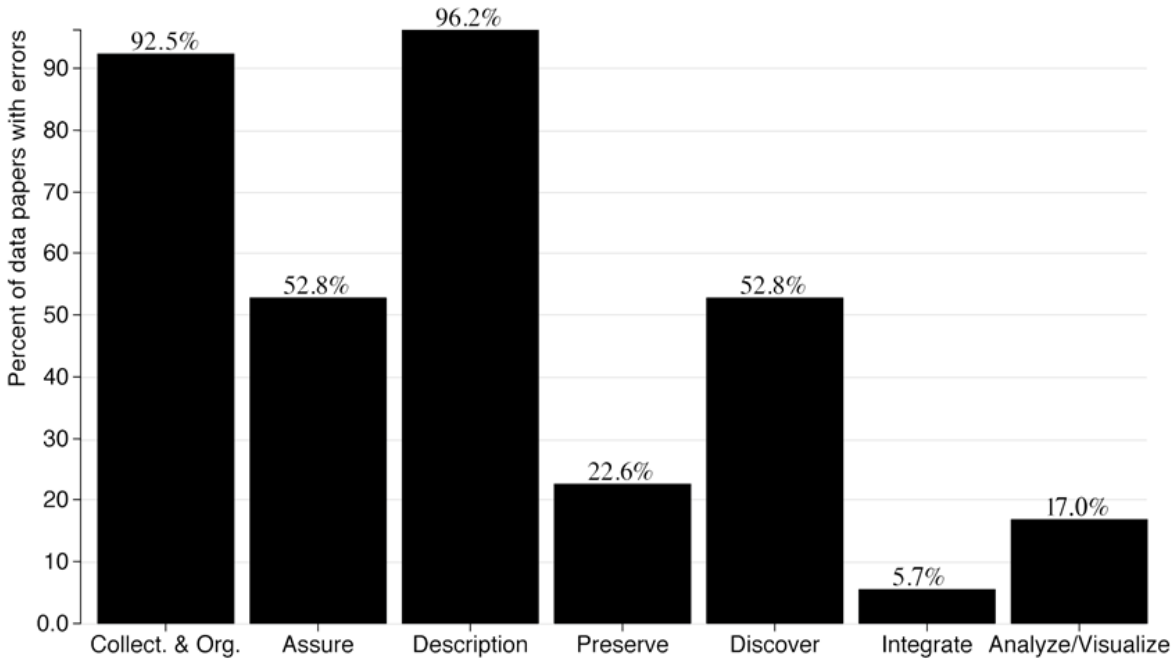
To maintain consistency, a single researcher (Author #1) performed all initial coding and classification of reviewer-identified errors. Occasionally, multiple reviewers pointed out the same error in a data paper. When this occurred, the error was counted once in the quantitative tally of errors performed later in the analysis process. This process resulted in the identification of more than 100 detailed categories, many of which were closely related. To narrow this down, overlapping categories and categories that contained only one or two identified errors were combined as appropriate, which resulted in 60 detailed error categories. Finally, the detailed error categories were grouped into Error Classes under the Data Life Cycle elements, where appropriate. After this process was complete, each overarching Data Life Cycle Element category had zero to four major error classes.

Quantitative analysis consisted of generating descriptive statistics related to each of the data life cycle elements and error classes. Initially, this entailed noting the total number of errors for each detailed category and Data Life Cycle element. This information was used to calculate the mean number of errors for the overarching Data Life Cycle elements, error classes, and detailed error categories. In the next step, the number of papers with each detailed error category was calculated, as well as the percent of papers with each error. Finally, the number of errors for each paper in the overarching Data Life Cycle elements and error classes was tallied. This allowed the calculation of the median number of errors for each category, as well as the mean and median number of errors per paper. Results of this quantitative analysis are presented below.

Results

Reviewers identified an average of 20.3 errors per data paper. The numbers of errors identified by reviewers varied yearly and there were no consistent long-term trends

Figure 2: Percent of data papers with errors in each Data Life Cycle Element Category. Each paper may have errors in multiple Data Life Cycle Categories.



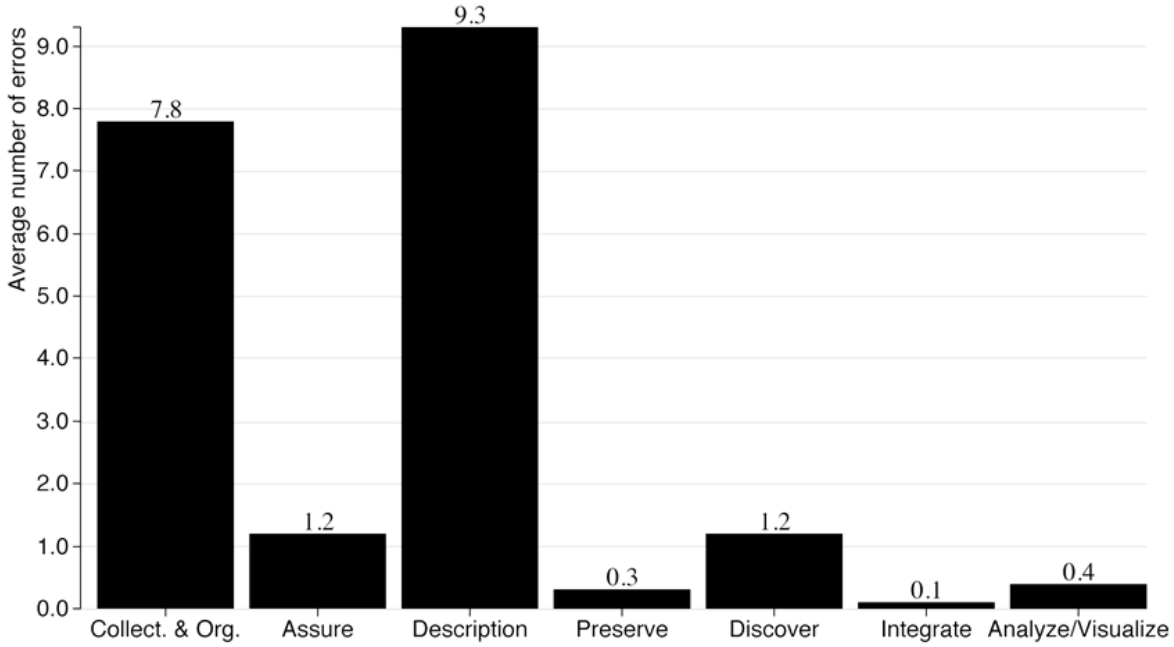
with respect to the overall number of errors (Figure 1). Through all years, the most common errors identified by reviewers were Collection & Organization and Description errors. Reviewers also consistently identified Assure errors each year, although these were significantly less common than Collection & Organization or Description errors.

Most data papers (49 out of 53; 92.5%) had errors associated with Collection and Organization (Figure 2). On average, each data paper had 7.8 Collection and Organization errors (Figure 3). The most common Collection and Organization errors were in the description of collection methods (38; 71.7%); not adequately describing the data collection site or time frame (29; 54.7%); and omitting relevant variables that were important for future analysis of the data set (43.4%). Nearly half (26; 49.1%) of the papers had an error in the description of the data collection protocol, including errors of omission, such as neglecting to explain how long samples, such as water or soil samples, were stored before analysis. Errors in the description of the data collection site included not describing how the site was determined or subdivided, including whether critical points of plots, such as edges or center points, were clearly marked.

Over half of the papers analyzed (28; 52.8%) had errors in the description of Quality Assurance/Quality Control (QA/QC) procedures (Figure 2), with an average of 1.2 errors per paper (Figure 3). Nearly one third of the papers (32.1%) did not adequately describe their QA/QC procedures. Errors ranged from neglecting to provide basic statistics regarding the data, such as ranges or mean values, to incomplete descriptions of logical consistency checks or benchmarks used to verify the accuracy of the data.

The most common Data Life Cycle Element errors were Description errors (51; 96.2%) (Figure 2), and data papers contained an average of 9.3 Description errors (Figure 3). Many such errors (83.0%) were simple edit-

Figure 3: Mean number of errors in a given Data Life Cycle Element Category.



ing errors, including grammatical errors (44; 58.5%) that ranged from awkward sentence structure or wordiness, to simple mistakes that an automatic grammar check would catch, such as missing spaces after a period. Errors in descriptive metadata were also very common (39; 73.6%) and many researchers (24; 45.3%) had a tendency to use either vague terms, such as "moderate" or "extreme," or field jargon, such as "degree of fragmentation," without clearly defining those terms. Finally, 43.4% (23) of the papers did not adequately describe the overall research project, such as not providing the background information required to get a clear understanding of the scientific context or questions that framed the study.

Reviewers of data papers noted errors related to the long-term preservation and storage of the submitted data sets in about one in five papers (12; 22.6%), (Figure 2). For example, an author may have mentioned that he or she kept all original data and records in personal offices or computers, or was storing data in proprietary or non-archival formats. Authors might not provide details regarding the maintenance of the data set, in cases of data sets archived over extended periods.

Over half of the papers (28; 52.8%) had Discover errors that would affect the ability to discover a particular data set and to assess the data set's utility (Figure 2). Despite the large number of papers with Discover errors, the average number of Discover errors per paper was much lower, with only 1.2 errors per paper (Figure 3). The most common Discover errors were insufficient description of access or use constraints (7; 13.2%); insufficient description of the data set's contributions and limitations (18; 34.0%); and not including information that would make finding the data set easier for potential data reusers (11; 20.8%). Of this last category, 17.0% were a result of authors not including all relevant information in the abstract such as not including the years of data collection or not summarizing the data collection methods.

About six percent (3; 5.7%) of these papers had errors in the integration of data sets

Table 2: Descriptive statistics for the most common error categories for each stage of the Data Life Cycle. These will not sum up to 100%, since each data paper may have multiple errors in any given Data Life Cycle category. The total number of papers analyzed was 53.

Data Life Cycle Element Categories/ Data Life Cycle Elements Sub-Classes/ Detailed Error Category	% (number) of Data Papers with error	Mean Number of Errors/ Paper	Median Number of Errors/ Paper
Collection & Organization	92.5% (49)	7.8	7
Collection methods	71.7% (38)	3.2	2
Protocol description	49.1% (26)	0.9	
Measure collection context	34.0% (18)	0.4	
Collection site / time description	54.7% (29)	1.6	1
Site description	43.4% (23)	0.8	
File structure organization	32.1% (17)	0.7	0
Fields combined	17.0% (9)	0.3	
Data presentation	73.6% (39)	2.3	2
Include all relevant variables	43.4% (23)	0.8	
Assure	52.8% (28)	1.2	1
QA/QC procedure description	32.1% (17)	0.5	
Description	96.2% (51)	9.3	6
Editing	83% (44)	5.2	3
Bibliographic	47.2% (25)	1.0	
Grammatical	58.5% (31)	3.6	
Metadata	73.6% (39)	3.3	2
Data dictionary	37.7% (20)	1.0	
Use of vague terms or jargon	45.3% (24)	1.0	
Study description	43.4% (23)	0.8	0
Background information	32.1% (17)	0.6	
Preserve	22.6% (12)	0.3	0
Non-computer readable formats	9.4% (5)	0.1	0
Discover	52.8% (28)	1.2	1
Constraints	13.2% (7)	0.1	0
Insufficient access description	9.4% (5)	0.1	0
Uses	34.0% (18)	0.6	0
Contributions and limitations	30.2% (16)	0.4	
Finding data set	20.8% (11)	0.4	0
Incomplete abstract	17.0% (9)	0.3	
Integrate	5.7% (3)	0.1	0
Failure to cite data sources	5.7% (3)	0.1	0
Analyze / Visualize	17.0% (9)	0.4	0
Description of analysis procedure	13.2% (7)	0.2	0

(Table 2). Each data paper that had an error of this type failed to properly cite the sources of data that went into the integrated data set. For example, one data paper provided climatic data to supplement the data collected, but neglected to acknowledge the source of the climatic data. Another data paper did not use the most current version of the referenced data source.

While most data papers presented raw data sets, numerous papers included some analysis of the data. Seventeen percent (9) of the data papers analyzed had some type of error in the presentation of the analysis or visualization results (Table 2). These errors included neglecting to include statistical significance of the analysis results, not including all relevant variables, and not explaining how the data changed during the analysis process. Of the nine papers that had Analyze/ Visualize errors, seven authors did not sufficiently describe their analysis methods, such as not documenting formulas used to create new variables or data sets.

Discussion

Data are an important product of research. Data to be re-used in the future requires the careful preparation of metadata and documentation that allows future users to find and understand it. In this case study, common errors observed from reviews of *Ecological Archives* were compiled and described; these errors serve as the basis for informing data documentation. Despite any limitations associated with focusing on ecological data, many of the errors identified are representative of those occurring in other fields of research. The analysis of the cause of these errors, along with existing data management practices (Michener et al., 1997, Cook et al., 2001, Borer et al., 2009, Hook et al., 2010) provide examples across research disciplines for data documentation and preparation.

Data Publication Best Practices for Researchers

Data collection and organization best practices

Researchers could avoid many errors by taking detailed notes before, during, and after the data collection process. This starts with describing the study and the goals for the study. Authors of nearly half the papers analyzed (43.4%) did not sufficiently describe the project background, goals, or research questions. This information is essential, since it describes the larger research project and provides the scientific context that shapes the decisions made regarding data collection and analysis (Strasser et al. 2012). Contextual information includes the spatial location of the data collection site, the time frame when data collection occurred, and environmental factors that could affect the observations and subsequent interpretation of the data. Photos, maps, and GPS coordinates of the data collection site are critical to data reuse, especially if future researchers choose to resample the area. This is especially important, since many sites are changing due to natural or human-caused changes.

Metadata associated with most (71.7%) data papers lacked sufficient detail about data collection process and methods, including experimental manipulations, measurements, and sampling choices made during the data collection process. Information about sampling designs, research methods, and identification of project personnel is central to interpreting and using data (Michener et al. 1997).

Data quality assurance and control best practices

Metadata from most data papers (52.8%) did not describe quality assurance and quality control (QA/QC) procedures in detail. Detailed descriptions of QA/QC procedures are critical for those looking to determine fitness

for use of the data (Hook et al. 2010). Without a complete record of all steps taken to ensure that research personnel measure and record the data correctly, researchers seeking to reuse data cannot accurately assess the quality of that data. Similarly, researchers who reuse data later need clear descriptions of any procedures performed to verify the data was accurately transferred (e.g., dual data entry—the comparison of data sets entered by two different individuals).

Metadata and description best practices

If researchers have taken complete notes throughout the data gathering and analysis process, then the task of creating metadata becomes much easier. Researchers can save time and effort by carefully considering what metadata will be necessary at the outset of the project, rather than trying to correctly recall important details after the data collection and analysis process is complete. For instance, creating and maintaining a data dictionary is easiest when done at the inception of the study, instead of waiting until the data are ready to publish. Similarly, fully defining all codes and variables, including measurement method, units of measurement, as well as field site names (if appropriate) is easiest when done at the time of data collection and analysis. When writing metadata, another helpful guideline for researchers to follow is to use the same rules they would use when writing a paper for publication. This includes defining any jargon or acronyms and running the descriptive metadata through a grammar and spell-checking tool prior to submission.

Data preservation best practices

Preserving data for future research requires careful consideration. One important aspect of this is to save data in a non-proprietary format, such as ASCII text, while avoiding saving data in non-extractable formats, such as PDF. Storing data in multiple places helps to protect data from accidental loss, even if researchers intend to publish a complete set of their data in a repository. If the original data is only stored on a personal computer or in an office file cabinet, it is especially prone to loss through accidents (Michener et al. 1997).

Data discovery best practices

When researchers publish data in a repository, the ancillary information included with that data is essential for other researchers who need to find the data later. This entails including all relevant information in the abstract and listing all keywords that could describe the data.

Data integration and analysis best practices

Clear descriptions of all data integration and analysis steps, including any software used to process the data, is just as important for those reusing data as understanding the methods used to collect the data in the first place. Documenting any changes to the data set is a key part of maintaining data provenance (Strasser et al. 2012) and is critical to enabling data re-users to assess the confidence that they can place on the data (Chapman & Jagadish 2007). One way to document provenance is to include scientific workflows and code from software scripts such as R in the metadata since they can provide a record of changes made to the data (Borer et al. 2009).

Various research support organizations such as DataONE (www.dataone.org), the Federation of Earth Science Information Partners (www.esipfed.org), and the Inter-university Consortium for Political and Social Research (www.icpsr.umich.edu) offer data management training and educational tools to specific communities of researchers. In addition, libraries and data librarians understand the importance of data management in 21st Century research and are increasingly providing critical research data management services (Tenopir et al. 2012). Because of their central and trusted roles in academia

and research, libraries and librarians may be in a prime position to directly or indirectly support data management education for faculty and students (Treloar et al. 2012). Existing training and education tools as well as the best practices documented herein can provide the foundation for improving data stewardship in the sciences.

Conclusions

This paper provides a case study of common and representative errors observed when researchers prepare data and documentation for sharing and archiving. The findings were derived from *Ecological Archives* but are also applicable for other research disciplines that require data management for long-term archive.

One objective of this paper was to identify common errors in data organization and metadata completeness that would preclude a "reader" from being able to interpret and re-use the data. Publishing data for sharing and reuse is error-prone and each stage of the data life cycle presents opportunities for mistakes. In the data collection stage, researchers failed to describe their methods, the data collection site, or the context in which the samples were collected. Errors in the QA/QC stage of the life cycle occurred when researchers did not describe validation procedures, either during data collection or data entry. The most common errors are those where the researcher did not provide metadata that was adequate to enable others to interpret and potentially re-use the data.

The second objective was to use these common errors to develop a set of best practices for data management that would guide researchers across disciplines in creating more usable data products. A set of recommendations for best practices for data publication, summarized by elements of the data life cycle, are presented to enable researchers from many disciplines to create data products that are easier to share and re-use.

References

Borer, Elizabeth T., Eric W. Seabloom, Matthew B. Jones, and Mark Schildhauer. "Some Simple Guidelines for Effective Data Management." *Bulletin of the Ecological Society of America* 90, no. 2 (2009): 205-214, http://dx.doi.org/10.1890/0012-9623-90.2.205

Chapman, Adriane, and Hosagrahar V. Jagadish. "Issues in Building Practical Provenance Systems." *IEEE Data Engineering Bulletin* 32, 4. (2007): 38-43.

Cook, Robert B., Richard J. Olson, Paul Kanciruk, and Leslie A. Hook. "Best Practices for Preparing Ecological and Ground-Based Data Sets to Share and Archive." Bul*letin of the Ecological Society of America* 82, no. 2 (2001): 138-141.

"Data Paper Instructions – ESA's Ecological Archives," last modified June 14, 2012, http://esapubs.org/archive/instruct_d.htm.

Hook, Les A., Suresh K. Santhana Vannan, Tammy W. Beaty, Robert B. Cook, and Bruce E. Wilson. "Best Practices for Preparing Environmental Data Sets to Share and Archive." Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A., (2010), http://dx.doi.org/10.3334/ORNLDAAC/BestPractices-2010. Retrieved May 27, 2012 from http://daac.ornl.gov/PI/BestPractices-2010.pdf.

Michener, William K., James W. Brunt, John J. Helly, Thomas B. Kirchner, and Susan G. Stafford. "Nongeospatial Metadata for the Ecological Sciences." *Ecological Applications* 7, no. 1 (1997): 330-342.

Michener, William K., and Matthew B. Jones, "Ecoinformatics: Supporting Ecology as a Data-Intensive Science." *Trends in Ecology & Evolution* 27, no. 2 (2012): 85-93. http://dx.doi.org/10.1016/j.tree.2011.11.016.

Moore, Allen J., Mark A. McPeek, Mark D. Rausher, Loren Rieseberg, and Michael C. Whitlock, "The Need for Archiving Data in Evolutionary Biology." *Journal of Evolutionary Biology* 23, no. 4 (2010): 659-660, http://dx.doi.org/10.1111/j.1420-9101.2010.01937.x

National Institutes of Health (NIH). "NIH Data Sharing Policy and Implementation Guidance." Last modified March 5, 2003. http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm.

National Science Foundation (NSF). "Grant Proposal Guide Chapter II.C.2." Last modified January 2011. http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp.

Rausher, Mark D., Mark A. McPeek, Allen J. Moore, Loren Rieseberg, and Michael C. Whitlock. "Data Archiving." *Evolution* 64, no. 3 (2010): 603-604, http://dx.doi.org/10.1111/j.1558-5646.2009.00940.x

Treloar, Andrew, G. Sayeed Choudhury, and William Michener. "Contrasting National Research Data Strategies: Australia and the United States." In *Managing Research Data*, edited by Graham Pryor, 173-203. London: Facet Publishing, 2012.

U.S. Government Accountability Office. 2007. "Climate Change Research: Agencies Have Data-Sharing Policies but Could Do More to Enhance the Availability of Data from Federally Funded Research." GAO-07-1172. http://www.gao.gov/products/GAO-07-1172.

U.S. Office of Science and Technology Policy. 2013. "Increasing Access to the Results of Federally Funded Scientific Research: Memorandum for the Heads of Executive Departments and Agencies." http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Rieseberg, Loren, Tim Vines, and Nolan Kane. "Editorial and Retrospective 2010." *Molecular Ecology* 19, no. 1 (2010): 1-22, http://dx.doi.org/10.1111/j.1365-294X.2009.04450.x

Strasser, Carly, Robert B. Cook, William K. Michener, and Amber Budden, "Primer on Data Management: What you always wanted to know," A DataONE publication available via California Digital Library (2012), http://dx.doi.org/10.5060/D2251G48

Tenopir, Carol, Robert Sandusky, Suzie Allard, and Ben Birch. "Academic Librarians and Data Research Services: Preparation and Attitudes." Paper presentation at the IFLA World Library and Information Congress, Helsinki, Finland, August 11-16, 2012.

Wenburg, John K. "Data Archiving." *Journal of Fish and Wildlife Management* 2, no. 1 (2011): 1-2, http://dx.doi.org/10.3996/1944-687X-2.1.1

Whitlock, Michael C., Mark A. McPeek, Mark D. Rausher, Loren Rieseberg, and Allen J. Moore, "Data Archiving." *American Naturalist* 175, no. 2 (2010): 145-146, http://dx.doi.org/10.1086/650340

Zhang, Yan, and Barbara M. Wildemuth. "Qualitative Analysis of Content," In *Applications of Social Science Research Methods to Questions in Library and Information Science*, edited by Barbara M. Wildemuth, 308-319. Englewood, CO: Libraries Unlimited, 2009.

Disclosure: The authors report no conflicts of interest.

ISSN 2161-3974

Appendix A: Complete error categorization.

Data Life Cycle Element Category and Sub-Class	Detailed Error Categories	% of papers
Collection & Organization		
Collection methods	Describe limitations of research design	30.2%
	Describe method protocol and how it was developed	49.1%
	Describe manipulations and effect of manipulations	13.2%
	Describe measurement procedures, including instrumentation	28.3%
	Measure all aspects of collection context	34.0%
	Describe sampling choices	24.5%
	Describe personnel who performed data collection	7.6%
	Describe permits, laws, or standards that affect data collection	2.0%
Collection site / time description	Describe all site attributes, including coordinates, bounding areas, and landscape features	43.4%
	Describe time frame of data collection, including season	24.5%
	Include site photos, diagrams, and maps	9.4%
File structure and organization	Fields combined when they should be separate	17.0%
	Lack of database relational key	13.2%
	Data spread across too many files	7.6%
	Data file structure does not allow for automatic processing	11.3%
	Include size of data files, both individual and total	7.6%
Data presentation	Use descriptive and unique labels (e.g. columns, codes, variables, etc.)	20.8%
	Provide units for all measurements	15.1%
	Include all relevant variables	43.4%
	Describe how missing data is represented	24.5%
	Use standard practices for entering data	24.5%
	Provide realistic accuracy	17.0%
	Do not use indices or calculated values without including raw values	13.2%
Assure		
	Describe all QA/QC procedures	32.1%
	Describe how taxonomic misclassification was avoided	7.6%
	Describe any benchmarks used	5.7%
	Describe any anomalous data	18.9%
	Provide basic statistics (e.g. ranges, median, quartiles, etc.)	17.0%

Appendix A (continued): Complete error categorization.

Data Life Cycle Element Category and Sub-Class	Detailed Error Categories	% of papers
Description		
Editing	Bibliographic entry errors	47.2%
	Grammatical errors	58.5%
	General spelling errors	28.3%
	Taxonomic errors	9.4%
Metadata	Include tables, figures, and visualizations	24.5%
	Include clear and accurate data dictionary	37.7%
	Use terms consistently	11.3%
	Avoid using vague terms or jargon	45.3%
	Metadata should be machine readable	3.8%
	Include metadata independent of web links	9.4%
	Metadata should be accurate and meet community standards	15.1%
	Describe any software used	5.7%
Study description	Describe background information regarding study	32.1%
	Describe project goals and objectives	13.2%
	Describe hypotheses and research questions	3.8%
	State funding sources	1.9%
Preserve		
	Describe maintenance of data set	7.6%
	Avoid proprietary and non-extractable formats	9.4%
	Data sets should be archived some place other than just a personal computer or office	7.6%
Discover		
Constraints	Describe any access constraints	9.4%
	Describe any use constraints	3.8%
Uses	Describe potential uses for data	13.2%
	Describe long term value of data	7.6%
	Describe scientific contributions of data	30.2%
Finding data set	Include all relevant information in abstract	17.0%
	Include all relevant keywords	7.6%

Appendix A (continued): Complete error categorization.

Data Life Cycle Element Category and Sub-Class	Detailed Error Categories	% of papers
Integrate		
	Cite all data sources used in data set	5.7%
	Include all relevant variables from integrated data sets, or provide reasoning for why variables were excluded	1.9%
Analyze/ Visualize		
	Describe all analysis methods	13.2%
	Describe changes made to the data set during analysis	3.8%
	Include all raw variables	5.7%
	Include all relevant statistics of analysis (e.g. statistical significance)	7.6%