International Association of Scientific and Technological University Libraries, 31st Annual Conference, 2010

Data curation in avian ecology: a case study from both the scientist's and librarian's view

Eric Snajdr
Indiana University - Purdue University Indianapolis, esnajdr@iupui.edu

Data curation in avian ecology: a case study from both the scientist's and librarian's view

Eric Snajdr
Indiana University - Purdue University, Indianapolis, USA
esnajdr@iupui.edu

Abstract

This case study of a data curation project, which is currently in progress, demonstrates how a team of scientists has worked, in partnership with librarians, to plan to preserve their scientific output in an institutional repository. In addition, this case study offers a unique perspective. The author worked as one of the scientists in this particular research group for 10 years and is currently a science librarian working on this data curation project. As a result, the author has been an "insider" in discussions in both the scientist and librarian camps and provides viewpoints from both the scientist and librarian lenses.

The research group in this case study is the Ketterson/Nolan Research Group, a team of avian biologists in the Department of Biology Indiana University, Bloomington, Indiana. This research team has focused on the ecology, behavior, and physiology of a songbird, the dark-eyed junco. The research output from this group's long-term (thirty year) study on this single species of songbird has resulted in rich data sets of a variety of subjects (e.g. population demographics, behavioral observations, DNA records, and natural history).

The research group and librarians are working toward more than just the preservation of data, but also the preservation of accompanying descriptive documents that place this large body of work into historical and educational contexts. Described within this case study are preliminary issues that the scientists and librarians have worked through as they have moved to preserve the research output in the library's institutional repository.

Keywords
Long term research, bird biology, data curation, institutional repositories

Introduction

Upon retirement many scientists in academic institutions across the country discard notebooks, documentation in the lab, as well as data files. In many cases, all that remains of the research are the official forms of research output (e.g. scientific publications and reports). In some instances, the discarding of data and documentation appears to involve no major loss. For example, in many research labs, by the time the information has been reported in scientific publications, the data, by all appearances, have already been fully gleaned.

However, if, on the other hand, the data were preserved and made publically available, the opportunity would open for other researchers to view the data along side of the corresponding official research output. This would allow the potential for researchers worldwide to 1.) possibly gain a deeper understanding of a particular experiment or even an entire body of research, 2.) more fully understand the subtleties of how the research was done, 3.) be able to more completely verify findings, and 4.) more accurately replicate an experiment. The preservation and public availability of data sets would also 5.) make it possible that, in the future, other researchers could use the data in new ways. For example, the data may be incorporated into a study builds on the pre-existing dataset. Or, the data could be reanalyzed or reworked as scientific understandings or methodologies evolve. Perhaps new linkages could be made that

would otherwise be impossible if the data were lost.

Some disciplines (e.g. genomics) have a well-defined system for where their data may be placed/archived. Other disciplines have no such structure in place. Institutional repositories can fill this important role of providing a platform for data preservation (see Choudhury, G.S. 2008; Heidorn, 2008; Witt 2008). Academic librarians can serve as data curators by preserving research data produced by scientists at their respective colleges and universities. The data curation project described in this case study describes an effort, which is currently in progress, to preserve research data/output of two professors in the Department of Biology, Indiana University, Bloomington, Indiana, in the Indiana University's institutional repository.

Overview of the research group

The scientists in this case study are the Ketterson/Nolan Research Group, a team of avian biologists in the Department of Biology Indiana University, Bloomington, Indiana. This research group has had a consistent and productive research output. To date, the lab's work has resulted in over 100 publications in scientific journals. This group, headed by Ellen D. Ketterson and Val Nolan Jr., has focused much of their work on the ecology, behavior, evolution, and physiology of a single species of songbird, the dark-eyed junco. For more background on this research group see (http://www.indiana.edu/~kettlab/index.html).

The research output from this group's long-term (thirty year) study on a single free living population of dark-eyed juncos in southwest Virginia has resulted in rich data sets of a variety of subjects (e.g. population demographics, behavior, DNA records, and natural history). For an overview of this long term study see (Clotfelter et al., 2004; Ketterson et al., 2001; Ketterson et al., 1996; Ketterson and Nolan, 1992). A primary focus of this long-term project in avian biology has been the exploration of specific research questions involving hormones and behavior. However, because the group also systematically collected a wealth of population/natural history data on the study species, much of the data are still un-mined.

Goal of the data preservation project

The goal of this project was to preserve the Ketterson Nolan Research Group's research output and supporting documents in Indiana University's institutional Repository, Scholarworks. This research group's collection of materials is unique because of its large scope. The data span a study that has continued for 30 years and which is still in progress. Therefore, the resulting collection in Scholarworks will not be a static collection. Documents will be added over time as the research project continues.

The long-term study of the research group is, in itself, unique because it is unusually well documented in the form of images, a video documentary project (in progress) that is descriptive and educational in scope, and through written documentation describing details of each year and each phase of the project. The research group's data will be preserved in Scholarworks along with these supporting descriptive documents placing this large body of work into historical and educational contexts.

The author worked as a researcher in this research group for over ten years and is currently a science librarian working on this data preservation project. As a result, the author has been an "insider" in discussions in both the scientist and librarian camps and discusses issues from both the scientist and librarian viewpoints.

Determining what to preserve

During the initial planning stages, the researchers met with several of the Scholarworks librarians in order to discuss the idea of using the institutional repository as a platform to preserve the lab's research data. Upon deciding to pursue this endeavor, the researchers then compiled

an inventory of their research output and accompanying documentation that they would like to preserve. For each item, the listing included the name, brief description, current format (e.g. paper, or electronic, if electronic they also listed the software used to create the document or file type of the document), quantity (e.g. for paper files the volume or number of pages, file size for electronic), year span of the data, and general notes regarding each item.

The researchers used several criteria while considering what items to preserve. The first question they asked themselves was, "What items were of historical/contextual interest?" Certain documents were deemed of importance because they provided a context for the official research output (e.g. published papers) as well as various data files. For example, particular documents provided background details of the work, essentially describing how this long-term study was conducted (e.g. grant proposals, documents describing annual cycle of the project, monthly tasks, each field season's organization, goals, priorities, research team composition, each field season's instructions for field techniques, as well as protocols for laboratory work. Many of these documents serve to explicitly illustrate how this research group organized and conducted their work. The preservation of these items will allow other researchers to replicate this work or to use these supporting documents as a guide in order to design a similar research effort.

A second consideration in determining what to save was the question of, "What items have potential value for future research?" Within this category were 1.) unanalyzed data and 2.) data that might prove to be of use when linked with data from other researchers, and 3.) data that are likely to attain new value in the future as new methods and scientific understandings emerge.

This research group also aims to use Scholarworks in order to share data with collaborators as well as among members of their own research group. So, this was a third consideration of what items to mount in the institutional repository.

Categories of data/research output to be preserved

There were a wide variety of kinds of data/research output that the group wished to permanently archive. These items were placed under three broad categories.

Documents describing the background of the project

- Grant proposals (describing intentions)
- Documents describing Annual Cycle of the project - Each field season's organization, goals, priorities, team composition, protocols, and instructions (field techniques)
- Images
- Videos

Results (data)

- Data collected each year (e.g. population information, nesting data, territory maps, behavioral observations, DNA)
- Yearly summaries (e.g. population data, nesting success/predation rates, cowbird parasitism, etc…)
- Databases containing data across all years of the project

Output

- Formal output (published papers)
- Less formal output (presentations, posters)

Discussion

Several questions and concerns were raised by the researchers when they first considered the possibility of using Scholarworks as a platform on which to preserve their research output. For example, they asked "What other possibilities were available for data storage/management or

permanent data preservation?" and "Were there higher profile places?" As mentioned earlier, some disciplines have a well-established system for where their data may be placed/archived. The group investigated potential data archiving platforms for avian biology as well as for long-term studies in ecology. No existing ecology sites fit the mold of what the research group hoped to accomplish. While some bird specific data sharing sites existed (e.g. eBird of Cornell Lab of Ornithology and National Audubon Society), these sites did not currently have the structure or overarching purpose to accommodate the entire body of research output that this research group wished to preserve.

Related to the investigation of these non-library affiliated data sharing/preservation sites was another concern. Of primary importance to the scientists was, "After adding a document to the IR, did they still retain the rights to the item?" Also, "Could the data be in Scholarworks and another repository at the same time?" Upon receiving the answer of "yes" to both of these questions, the scientists did not have any reluctance in proceeding with the planning. They retained the rights to all items placed in the repository. Also, in the future, if a discipline specific data storage/sharing platform were created that suited their needs, they could place their data in that repository as well.

The institutional repository provides a secure, permanent site where data will be accessible to the world. However, because the Ketterson/Nolan Research Group still plans to analyze and publish some of the data that will be mounted in Scholarworks, the group will assign specific embargo periods on such items. In these cases, the data will be, for a specified span of time, accessible through Scholarworks only to the members of the research group and select collaborators. After the expiration of the embargo periods, these items will then be accessible to the world.

Related to this issue above, is the fact that the Ketterson/Nolan Research Group contains a mix of student researchers (post-doctoral, doctoral, and undergraduate). In addition to student involvement in the lab's research, there are several cases where an aspect of the research was performed in collaboration with researchers at other institutions. In many cases, the lead scientist has obligations to these students and collaborators with regard to data accumulated in a collective effort. Several of these students and collaborators may still have a stake in specific data files even after the lead researcher's retirement. Therefore, embargo periods need to be placed in order to protect the individual situations and people tied to each data set. This will ensure the specific parties involved, those with a stake in the data, have a chance to fully utilize any data before it becomes open to public.

Currently, the Ketterson/Nolan Research Group is working with the Scholarworks librarians to finalize the structure of the Ketterson/Nolan Scholarworks site and develop a final plan for the metadata that will accompany each item uploaded into the site. It is clear that quality metadata construction will be crucial in order to allow users to locate relevant items through search queries. Since many of the data files are numerical in nature, the ability for a user to locate relevant data files will rest solely on the metadata itself. However, there is a tradeoff in this process. Because of the high number of files the Ketterson/Nolan Research Group would like to preserve, this task could quickly become cumbersome. In other words, if too much time or effort is directed to constructing metadata, the progress of uploading and preserving the files will be quite slow.

A similar tradeoff exists with the annotation of data files. Certain data files of the research group include a wealth of information. However, in some cases, the information involves abbreviations or codes. The group already has several "data dictionaries" (files that describe the categories of data in detail). However, these were created for the research group themselves, not individuals unfamiliar with the research project. For a scientist outside of this research group, more extensive annotation will likely be needed. If the files are not understandable by users, the preservation of such files will likely be useless. So again, a tradeoff exists between needing to describe, but the more time much time that is taken in the description then the more time consuming it will be to adequately preserve the files in a way that will make them useful.

Importance of preserving long term data

It is interesting to note that the preservation of data is of interest to funding agencies as well. The National Science Foundation, which has provided the majority of funding for the Ketterson/Nolan Research Group's long term project, has become increasingly concerned with the issue of data preservation and data sharing. Below is an excerpt from instructions for writing a grant proposal for Long Term Research in Environmental Biology (National Science Foundation, 2010).

> "…proposals must include a section that clearly articulates how and where data will be archived, how they will be made available to the broader scientific community, and how future access to the data will be guaranteed."

In the case of this research group's work, the preservation of the data takes on a specific importance because it involves a long-term study (National Science Foundation, 2010).

> "Long-term data are a valuable resource that can stimulate and support investigations well beyond the scope of the initial study. The value of these data depends on their availability, along with the availability of associated metadata."

Long term research spanning decades is more the exception then the rule. However, long-term data sets are valuable, in one sense, because it is only through such data that certain scientific questions can be answered (National Science Foundation, 2010). The Ketterson/Nolan Research Group has already used a portion of their long-term bird population data (in this case the seasonal reproductive success of the songbird that they study) in collaboration with long term studies of other biologists in order to address questions of a broad ecological perspective. For example, one study (Clotfelter et al., 2007) spanned data collected from 1980 through 2004 and examined interactions among various plant and animal species in the same forest. This work would not have been possible without each of these scientists (which focused on completely different study species) sharing data that they systematically collected and preserved over multiple decades. In the future, it is likely that many other linkages such as those in this example could be made. And it is likely that many connections or uses of the data from the Ketterson/Nolan Research Group are beyond the imagination of scientists today.

Acknowledgements: The author expresses gratitude to The Library Fund, a fund of the Indianapolis Foundation, for the award of the Minde Browning Memorial Grant. The author also wishes to thank Ellen Ketterson, Nicki Gerlach, Garett Montanez, Julie Bobay, Jennifer Laherty. Sherri Michaels, Phillip Bantin, and Roger Beckman for their assistance with this project.

References

Choudhury, G.S. 2008. Case study in data curation at Johns Hopkins University. Library Trends, 57: 211–220.

Clotfelter, E.D., Pedersen, A.B., Cranford, J.A., Ram, N., Snajdr, E. A., Nolan Jr., V. and E.D. Ketterson. 2007. Acorn mast drives long-term dynamics of rodent and songbird populations. Oecologia 154: 493-503.

Clotfelter E.D., O'Neal, D.M., Gaudioso, J.M, Casto, JM, Parker-Renga, I.M., Snajdr, E.A., Duffy, D.L., Nolan, V. Jr., and E.D. Ketterson. 2004. Consequences of elevating plasma testosterone in females of a socially monogamous songbird: evidence of constraints on male evolution? Hormones and Behavior 46:171-178.

Heidorn, P.B. 2008. Shedding Light on the Dark Data in the Long Tail of Science

Library Trends, 57: 280-299.

Ketterson, E.D., Nolan, V. Jr., Casto, J.M., Buerkle, C.A.,Clotfelter, E., Grindstaff, J.L., Jones, K.J., Lipar, J.L., McNabb, F.M.A., Neudorf, D.L., Parker-Renga, I., Schoech, S.J., and E. Snajdr. 2001. Testosterone, phenotype, and fitness: a research program in evolutionary behavioral endocrinology. Pp. 19-40, in A. Dawson and C.M. Chaturvedi (eds.), Avian Endocrinology. Narosa Publishing House, New Delhi, India.

Ketterson, E.D., V. Nolan Jr., M.J. Cawthorn, P.G. Parker, and C. Ziegenfus. 1996. Phenotypic engineering: using hormones to explore the mechanistic and functional bases of phenotypic variation in nature. Ibis 138: 1-17.

Ketterson, E.D. and V. Nolan, Jr. 1992. Hormones and life histories: an integrative approach. American Naturalist 140:S33-S62.

National Science Foundation. 2010. Long Term Research in Environmental Biology (LTREB). Retrieved March 20, 2010 from http://www.nsf.gov/pubs/2010/nsf10558/nsf10558.htm

Witt, M. 2008. Institutional Repositories and Research Data Curation in a Distributed Environment. Library Trends, 57: 191-201.