A Residential Data Curation Internship: Opportunities and Challenges

Kelly M. Gordon [1]
Temporary Research Data Specialist
Albert R. Mann Library
Cornell University
Ithaca, NY 14853 USA

ABSTRACT

During the summer of 2008, while a student at San Jose State University's School of Library and Information Science, the author completed a residential data curation internship at the Cornell Biological Field Station. The internship entailed preparing and documenting a long term dataset collected by researchers at the field station. Internships such as this one present learning opportunities not only for students interested in the field of data curation, but for researchers as well. They also provide the opportunity to facilitate collaborations between library staff and researchers.

1. BACKGROUND

Recognizing researchers' need for data curation services early in the research cycle, librarians at Albert R. Mann Library at Cornell University applied for and received funding from the National Science Foundation to develop a local, institution-based data staging repository (DataStaR) for researchers affiliated with Cornell [1]. The intent of DataStaR is to provide researchers with the tools and support they need to perform data curation tasks in preparation for transmission of datasets to appropriate domain-based long-term data repositories. One goal of the project has been to foster partnerships with data owners in order to facilitate the development of this model.

One such partner is the Cornell Biological Field Station (CBFS), which serves as a primary field site for aquatic research at Cornell University. The centerpiece of the station's research program is a 50-year dataset on the food web of Oneida Lake, New York. Researchers at the field station, increasingly aware of the need to document and preserve this resource, saw the potential benefits of participation in the DataStaR project. Since the DataStaR platform and tools were not yet ready for use by researchers, DataStaR and CBFS personnel decided to hire an intern to carry out the work of preparing the dataset and creating metadata for transmission to a data repository.

2. INTERNSHIP

Mann Library and CBFS personnel collaborated in the hiring of an intern to complete a ten-week data curation internship during the summer of 2008. The successful candidate (the author of this paper) is an MLIS student enrolled in San Jose State University's School of Library and Information Science. Since the intern would need to work closely with CBFS researchers to prepare the dataset and develop metadata, it was agreed that she would live and work at the field station, but be remotely supervised by Gail Steinhart, one of the primary investigators on the DataStaR project. CBFS hosts a field research internship program for undergraduates, so the data curation intern was one of eight interns working at the field station that summer.

The long-term database used by researchers at CBFS consists primarily of multiple tables in Microsoft Access format. The intern prepared data and created metadata for seven data packages derived from this dataset. Metadata were created using the EML (Ecological Metadata Language) metadata standard [2]. Since the DataStaR platform is still a work in progress, the intern used the software package Morpho [3] to create EML records. The data packages were deposited in the Knowledge Network for Biocomplexity [4], one of the pre-eminent domain-based repositories for ecological datasets, and Cornell University's e-Commons [5], an institutional repository that houses electronic content produced by members of the Cornell Community.

3. OPPORTUNITIES

Residence at CBFS provided a number of opportunities for interactions above and beyond those related to the tasks of the internship. The intern's involvement with the day-to-day activities of the field station provided a deeper understanding of the issues faced by researchers prioritizing data management tasks in the context of other activities in the research cycle. The intern also participated in field data collection and discussions of ongoing research. One-on-one interactions with field station personnel in the course of carrying out internship tasks provided insight into the varying attitudes of researchers towards data management activities.

The presence of an intern actively involved in data curation tasks presented learning opportunities for station personnel, as well. Although most experienced researchers are well aware of the importance of data curation in a general sense and are eager to work to preserve and document their data, many lack an understanding of the nuts and bolts of data curation, and may not be aware of the role data repositories can play in data preservation. As the intern consulted with station personnel at

[1] Current author contact information: Bioscience and Natural Resources Library, University of California, Berkeley, Berkeley, CA94720.
+1 (510) 642-2030
kgordon@library.berkeley.edu

each step of the process, researchers gained a deeper understanding of the tasks involved in the preparation of data and metadata for transmission to a data repository.

Students, new to research, may not have previously considered the importance of data management, especially given the greater attention that data collection and analysis are likely to receive from their advisors. Living and working with graduate and undergraduate students at CBFS provided the opportunity for the data curation intern to engage in a number of informal dialogs about the potential pitfalls of data neglect, and about the benefits of adequate data documentation and preservation.

4. CHALLENGES

Data curation tasks are likely to be pushed to the back burner by researchers engrossed by managing data collection efforts, analyzing data, and writing publications. One challenge encountered during this internship was the need to frequently consult with field station personnel who were often at conferences or collecting data in the field. Researchers at the field station were sensitive to the fact that their input was needed to keep the project moving forward, but some aspects of the work were occasionally stalled by the unavailability of one or more key people. This issue was ameliorated by communication regarding researchers' and library staff schedules, and careful planning about the timeline on which specific tasks would be completed.

It is not uncommon for unanticipated issues to become evident when working with complex datasets that are handled by a number of different workers with varying degrees of familiarity with good data practices. Although some data preparation and cleanup tasks were planned for at the outset of this internship, it was necessary to assess each new issue as it arose to determine whether resolving it fell within the scope of the internship duties and whether it was feasible to attempt to correct the problem or if that responsibility would fall to the researcher. In general, when library personnel collaborate with researchers to perform data curation tasks, it may be desirable to define many data cleanup tasks as the responsibility of the researcher, in order to avoid an unsustainable drain on library personnel's time.

Even the most user-friendly, flexible platform for data curation tasks may present barriers to use if the platform can't be integrated into researchers' existing workflows [6]. Researchers adopting new data curation practices may find that the learning and use of tools require unanticipated changes to their current data habits. Additionally, if data management tasks have been a lower priority in the past, datasets may be in need of more extensive preparation than is feasible for researchers. The need to prepare data may present the most daunting barrier to preservation of data, yet librarians can provide limited assistance with these sorts of tasks, since the specifics are peculiar to each data set.

The presence of an intern working on data curation tasks acts as a spur to researchers to provide necessary information and support to keep the work moving forward. However, the flip side of this is that when the internship ends, data curation tasks may be deprioritized. Additionally, the work done during this particular internship represented an intensive level of data service that, although eagerly accepted by CBFS researchers, is ultimately impossible for library personnel to sustain. Nevertheless, CBFS researchers have gained some familiarity with the tasks involved in preparing the data sets themselves for publication, and have agreed to assume responsibility for this portion of the process, while Mann Library staff continue to provide support for the creation of metadata. As the DataStaR platform is more fully developed and tools become available, CBFS personnel will assume greater responsibility for the entire publication process, with the librarian serving in an advisory capacity.

5. CONCLUSIONS

Graduate student internships can provide an excellent vehicle for the promotion of collaboration between researchers and librarians in accomplishing data curation goals. In particular, internships that embed the intern in the daily working environment of the researcher not only provide vital opportunities for interaction related to the task at hand, but also allow both researchers and the intern to benefit from an enhanced understanding of the particulars of data curation tasks, on the one hand, and of the research environment, on the other. Excellent communication regarding researchers' schedules and the data services that will and will not be provided by the intern and other library personnel can help to smooth over some of the challenges that may be encountered during a collaboration such as this one.

6. ACKNOWLEDGMENTS

Sincere thanks to Gail Steinhart of Cornell University's Albert R. Mann Library, and to Ed Mills, Lars Rudstam, Kristen Holeck, Tom Brooking, and the staff of Cornell Biological Field Station for their invaluable assistance. This work is supported by National Science Foundation grant number III-0712989.

7. REFERENCES

[1] Cornell University Library. 2007. About DataStaR. Online: http://datastar.mannlib.cornell.edu/about. Retrieved September 27, 2008.

[2] Ecological Metadata Language (EML). Online: http://knb.ecoinformatics.org/software/eml/. Retrieved September 27, 2008.

[3] Higgins, D., Berkley, C., & Jones, M. B. (2002). Managing heterogeneous ecological data using Morpho. *Proceedings of 14th International Conference on Scientific and Statistical Database Management, 24-26 July 2002,* 69-76. Online: http://dx.doi.org/10.1109/SSDM.2002.1029707 Retrieved September 27, 2008.

[4] Knowledge Network for Biocomplexity (KNB). Online: http://knb.ecoinformatics.org . Retrieved September 27, 2008.

[5] Cornell e-Commons. Online: http://ecommons.library.cornell.edu . Retrieved September 27, 2008.

[6] Steinhart, G. DataStaR: An Institutional Approach to Research Data Curation. *IASSIST Quarterly,* 2008, submitted.