How Information Science Professionals Add Value in a Scientific Research Center

Christopher Eaker
University of Tennessee
ceaker@utk.edu

Andrea Thomer
University of Illinois Urbana-Champaign
athomer2@illinois.edu

Erica Johns
University of Tennessee
ejohns3@utk.edu

Kayla Siddell
University of Tennessee
ksiddell@utk.edu

Abstract

In response to the increasing need for a data curation workforce, the Data Curation Education in Research Centers program is educating library and information science students in scientific data curation. During the summer of 2012, the authors worked alongside scientists and data managers at the National Center for Atmospheric Research in Boulder, Colorado, to learn data curation within the context of a research center. Each student was matched with a "Science Mentor" and a "Data Mentor" based on prior work experience and the results of a placement questionnaire completed before the internship. Though NCAR has robust data services, we found that there was nevertheless still a role for data curators who can foster close collaborations between scientists and repository managers that may have traditionally not existed. This collaboration supports an otherwise impossible mutual education that benefits all involved. This poster demonstrates tangible outcomes of these internships.

Keywords: data curation, data management, metadata, data citation, data life cycle

Introduction

The volume of scientific data is growing exponentially across all scientific disciplines (Lynch, 2008; Hey & Trefethen, 2003). In response to the increasing need for a data curation workforce, the Data Curation Education in Research Centers (DCERC) program is educating library and information science students in scientific data curation through a partnership between the iSchools at the University of Illinois (Illinois) and the University of Tennessee (UT) and the National Center for Atmospheric Research (NCAR) (Palmer, Allard & Marlino, 2011). The project, funded by the Institute for Museum and Library Services, is supporting three master's students from UT and three doctoral students from Illinois with diverse backgrounds in science and research. This poster reports on the summer internships at NCAR, a critical part of the DCERC program. It will describe the student projects, in particular, their role supporting collaboration and the phases of the data life cycle (Figure 1), as well as display visual artifacts produced during the projects.

Background

During the summer of 2012, four students (three from UT and one from Illinois) worked alongside scientists and data managers at the National Center for Atmospheric Research (NCAR) in Boulder, Colorado, to learn data curation within the context of a research center. NCAR has long-standing, sophisticated data services (i.e., extensive IT support staff, data repositories and repository managers), providing an ideal environment for students to learn current best practices in data management in a state-of-the-art research environment. The students brought LIS perspectives and expertise in areas including information organization, user communities, and long-term preservation to the internship teams at NCAR (Palmer, Renear & Cragin, 2008), as well as skills obtained from prior careers. Their projects demonstrated the valuable roles that LIS data curation can bring to a large-scale data center setting, especially as intermediaries who foster collaboration between all project stakeholders.

Eaker, C., Thomer, A., Johns, E., & Siddell, K. (2013). How information science professionals add value in a scientific research center. *iConference 2013 Proceedings* (pp. 812-815). doi:10.9776/13397
Copyright is held by the authors.

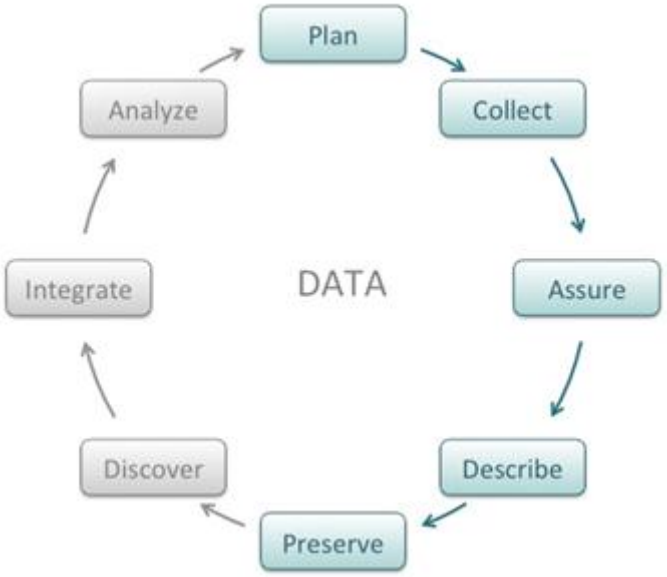

Figure 1. DataONE Data Life Cycle. From "Participatory design of dataONE -- Enabling cyberinfrastructure for the biological and environmental sciences," by W. K. Michener, S. Allard, A. Budden, R. B. Cook, K. Douglass, M. Frame, S. Kelling, R. Koskela, C. Tenopir, and D. A. Vieglais, *Ecological Informatics*, 11(0), pp. 5-15. doi:10.1016/j.ecoinf.2011.08.007. Copyright W. K. Michener. Reprinted with permission.

The Projects

Each student was matched with a "Science Mentor" and a "Data Mentor" based on prior work experience and the results of a placement questionnaire completed before the internship.

Chris Eaker (Data Mentor, Scot Loehrer; Science Mentor, Kate Young). The project addressed the following areas of the data life cycle: Collect, Describe, Preserve, and Analyze. With guidance from his mentors, Eaker conducted a data management audit within the In-situ Sensing Facility (ISF) in NCAR's Earth Observing Laboratory (EOL). The audit identified areas of strength, areas needing improvement, redundant activities, and useful tools within groups' data management practices. The goal of the audit was to increase efficiency and improve data management practices within the research groups, which will, in turn, support the long-term preservation and access of data sets. He produced a report with recommendations for improvements to ISF's data management workflows in areas including data management planning, metadata, archiving and preservation, and project tracking.

Erica Johns (Data Mentor, Robert Dattore; Science Mentor, Dr. Samuel Levis). The project addressed the following phases of the data life cycle: Plan, Collect, Assure, Describe, Preserve, Discover, and Integrate. With guidance from her mentors, Johns accumulated, reformatted, and ingested scientific data from the Ameriflux network for an NCAR climate scientist, all the while seeking to create a valuable data product that would be accessible for future users through NCAR's Computational and Information Systems Laboratory (CISL) Research Data Archive (RDA). (http://bit.ly/PH8STV). Johns then compared the data curation life cycle as she experienced it to the published DataONE model (Figure 1).

Kayla Siddell (Data Mentor, Steve Worley; Science Mentor, Dr. Patricia Romero-Lankao). The project addressed the following phases of the data life cycle: Plan, Collect, Assure, Describe and Preserve. Siddell acted as a translator and facilitator between a social scientist and repository manager by curating the first cross-disciplinary data set into the RDA. The final deliverable was a curated and archived data set, accessible on the RDA website (http://bit.ly/S6gvZf) and a report on the lessons learned over the seven week internship containing data management tips for the research team.

Andrea Thomer (Data Mentor, Gary Strand; Science Mentor, Dr. David Schneider). The project addressed the following phases of the data life cycle: Describe, Discover, and Integrate. Thomer and her

mentors broadly explored ways of fostering a culture of metadata creation and data citation among researchers, and specifically explored ways of incentivizing contributions to the Climate Data Guide (CDG) – the "go-to source for scientifically sound information and advice on the strengths, limitations, and applications of climate data" (**climatedataguide.ucar.edu**). They added easily copy-and-pasteable "suggested citation" text to the top of each page in the CDG, in which contributors to the guide are listed as authors. They hope that by making the CDG a more formal, citable publication, scientists will be encouraged to make contributions. Additionally, they looked at metadata use in climate model output data to assess how often researchers made use of existing metadata standards. A report was written containing their findings.

Discussion

Though all four of these projects focused on different aspects of data curation and different portions in the data curation life cycle, they were similar in that the interns consistently acted as intermediaries between different groups in one of the NCAR research centers. Siddell, for instance, acted literally as a translator between the scientist and the repository manager, working closely with Romero-Lankao to first understand her data set, which was often written in Spanish and, in some cases, Spanish abbreviations, and then with Worley to understand how to best prepare the data set for ingest into the RDA. Similarly, Eaker's project demonstrated the value of collaboration among researchers within the different research groups in the same laboratory. His project aimed to bring together researchers from separate groups to share tools, data management practices, and metadata to improve efficiency and aid in long-term preservation of data sets. Finally, Johns' project required continuous collaboration with the scientists using Ameriflux data, the data provider, and the software engineer who made the final reformat and ingestion possible. She repeatedly appraised the usefulness and appropriateness of the data obtained for integration into one data product that would not only meet the current scientists' needs, but would also be a valuable data product for future users.

In addition to serving as intermediaries, Information Science professionals trained as data curators are particularly knowledgeable about metadata and can assist researchers with describing their research data completely. Metadata allow scientists to make sense of data, i.e., how they were collected, why they were collected, who collected them, and how they were processed (Michener & Jones, 2012). Thomer found that scientists needed to be incentivized to add appropriate metadata to their data sets at multiple points in the data life cycle -- not just at the point of data creation, but also after first use. Eaker's project confirmed what Tenopir, et al., discovered: researchers often do not use a metadata standard at all or use one that is specific to their laboratory (2011).

Conclusion

Though NCAR has robust data services, we found that there was nevertheless still a role for data curators who can foster close collaborations between scientists and repository managers that may have traditionally not existed. This close collaboration supports an otherwise impossible mutual education. The internships helped data managers on the team learn more about user needs, and researchers on the team gain a better and stronger understanding of the importance of good data management. Furthermore, in addition to the tangible outcomes of the students' projects, the scientific researchers benefit by gaining an understanding of the importance of metadata at all steps during their project.

References

Hey, T., & Trefethen, A. (2003). The data deluge: An e-science perspective. *Grid Computing* (pp. 809-824): John Wiley & Sons, Ltd.

Lynch, C. (2008). Big data: How do your data grow? *Nature*, 455(7209), 28-29.

Michener, W. K., Allard, S., Budden, A., Cook, R. B., Douglass, K., Frame, M., . . . Vieglais, D. A. (2012). Participatory design of dataONE -- Enabling cyberinfrastructure for the biological and environmental sciences. *Ecological Informatics*, 11(0), 5-15. doi:10.1016/j.ecoinf.2011.08.007

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology & Evolution*, 27(2), 85-93. doi:10.1016/j.tree.2011.11.016

Palmer, C., Allard, S., & Marlino, M. (2011). Data curation education in research centers. *Proceedings of the 2011 iConference*, 738–740. Retrieved from http://dl.acm.org/citation.cfm?id=1940891

Palmer, C., Renear, A., & Cragin, M. (2008). Purposeful curation: Research and education for a future with working data. *Proceedings of the 2008 IDCC conference*, (1972), 1986. Retrieved from http://www.ideals.illinois.edu/handle/2142/9764

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., . . . Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*, 6(6), e21101. doi:10.1371/journal.pone.0021101