MANAGING THE DATA DELUGE: UNDERSTANDING SCIENTISTS’ NEED FOR DATA CURATION SERVICES

JEANINE M. SCARAMOZZINO, COLLEGE OF SCIENCE AND MATHEMATICS LIBRARIAN
MARI R. RAMIREZ, DIGITAL REPOSITORY LIBRARIAN
KAREN MCGAUGHNEY, ASSISTANT PROFESSOR OF STATISTICS
CALIFORNIA POLYTECHNIC STATE UNIVERSITY - SAN LUIS OBISPO

Faculty Behaviors

Q. Select the two individuals or entities most responsible for data management during the research phase of your projects.
Q. Please select the two individuals or entities most responsible for data management after the research phase of your projects.

Back-up

Q. In general, are your digital data backed up?
Q. Please select the main location(s) where your active research data are typically backed up.
Q. Please select the main location(s) where your past research data are typically backed up.
Q. In general, how often are your digital data backed up?

Funding

Q. For my funded research, I include data management costs (i.e., hardware, software, technical support, etc.) in my research budget?
Q. I have a formal research data management plan for preserving my digital data.
Q. I migrate my data to new media as older technologies become obsolete.
Q. I reuse data for other projects (i.e., publications, posters, experiments, etc.).

Preservation

Q. I follow criteria for preserving my data.
Q. I have a formal research data management plan for preserving my data.

Metadata

Q. I create descriptive information (i.e., variable descriptions, metadata, data history, collection notes, etc.) about my data.

Re-use - By Others

Q. I share my data with others beyond my research group.
Q. I educate myself or best practices for preserving my data.

Sharing

Q. I educate myself or best practices for preserving my data.*

Faculty Attitudes

Responsible Party

Q. Who should be directly responsible for the long-term management of your digital data management during the research phase of your project(s)?
Q. Do you think it is important to back up your data?
Q. Where should your digital data be stored long-term?

Back-up

Q. Do you know what to do to ensure that the data are not accidentally modified or destroyed?
Q. Do you think it is important to back up your data?
Q. Where should your digital data be stored long-term?

Funding

Q. For your funded research, do you think it is important to include data management costs in your research budget?
Q. I have a formal research data management plan for preserving my digital data.
Q. I migrate my data to new media as older technologies become obsolete.
Q. I think it is important to openly share your data with others?
Q. Do you think it is important that your colleagues openly share their data?

Metadata

Q. Are you aware of criteria for the creation of descriptive information to aid in discovery and reuse of data?
Q. I have a formal research data management plan for preserving my digital data.
Q. I migrate my data to new media as older technologies become obsolete.

Re-use - By Others

Q. I would like more guidance on best practices for preserving your data.
Q. I would like more guidance on best practices for preserving your data.
Q. Are you confident in your data management skills?

SUMMARY

Data curation is defined as “…the active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education…activities that enable data discovery and retrieval, maintain its quality, add value, and provide for in-use over time, and this new field includes authentication, archiving, management, preservation, retrieval, and representation.”

While library research on data curation is active and ongoing in the humanities and social sciences, the research regarding data curation within the sciences is in its infancy. The lack of knowledge about data curation, management, and reuse has a direct impact on librarianship, library services, and library policies. Librarians are now being asked to provide services to archive data created at their universities. What are the data curation needs on campus, and what services are librarians and libraries willing and able to provide to meet these needs?

Information gathered from a survey distributed to Cal Poly State University San Luis Obispo science and mathematics faculty will help provide insight into the awareness of science researchers about data curation issues and their needs for data curation services and education regarding maintenance and management of data.

BACKGROUND

- Data curation -- the active and ongoing management of data created by scientific endeavors -- is the next big collection development and faculty education issue to face librarians.
- Academic publishers are beginning to require authors to submit datasets in concert with their completed manuscripts. Many funding agencies also require faculty to include a data management plan as a component of grant applications.
- Anecdotal evidence suggests that Cal Poly faculty do not actively curate their datasets after the completion of their research and some do so poorly during the research process. Based on these patterns, we believe faculty will need to be better informed on how to care for their data to sufficiently satisfy publisher and funding agency requirements.
- While many well-funded “R1” academic institutions (e.g., MIT, Purdue, Cornell) are actively engaged in curation of large datasets, little attention is being paid to small datasets. This is consistent with the research regarding data curation within the sciences in its infancy. Based on these patterns, we believe faculty will need to be better informed on how to care for their data to sufficiently satisfy publisher and funding agency requirements.
- Many academic libraries have an institutional repository infrastructure in place to support the acquisition and delivery of published information.
- Academic publishers are beginning to require authors to submit datasets in concert with their completed manuscripts. Many funding agencies also require faculty to include a data management plan as a component of grant applications.
- Because many academic libraries have an institutional repository infrastructure in place to support the acquisition and delivery of published information, libraries are uniquely poised to provide technical support and education on metadata creation, retention, and migration of datasets.

PROJECT OBJECTIVES

- The aim of this research is to:
  - Understand scientists’ current data management activities.
  - Assess faculty awareness of data curation issues.
  - Identify gaps in scientists’ understanding of best practices for maintenance and management of data.
  - Identify educational opportunities to enhance researchers’ data management practices.
  - Develop reusable educational modules to educate faculty on data curation best practices.

METHODOLOGY

- Survey responses will be analyzed to determine faculty data curation behaviors and faculty attitudes.
- Responses will also reveal the level of awareness of data management best practices among faculty.
- Questions on current data management practices are generally paired with questions on faculty attitudes in order to highlight mismatches which may indicate a need for data curation services.
- If severe mismatches are found, it may indicate that faculty need help optimizing their data curation practices. The library can then consider appropriate strategies to provide such help.
- Based on the current awareness level, educational initiatives can be developed by the library to inform faculty of data curation issues and strategies.

SELECTED READING


ACKNOWLEDGEMENTS

- We would like to thank Anna Gold, Associate Library Services Dean, Michael Miller, Library Services Dean, Phil Bailey, College of Science and Mathematics Dean, and the COSAM Department Chair and faculty for their support of this research.
- We would also like to thank the COSAM faculty who provided valuable feedback on the draft survey.

ANALYSIS AND NEXT STEPS

Download this poster: http://digitalcommons.calpoly.edu/lib_fac/56