A Study of Faculty Data Curation Behaviors and Attitudes at a Teaching-Centered University

Jeanine Marie Scaramozzino, Marisa L. Ramírez, and Karen J. McGaughey

Academic libraries need reliable information on researcher data needs, data curation practices, and attitudes to identify and craft appropriate services that support outreach and teaching. This paper describes information gathered from a survey distributed to the College of Science and Mathematics faculty at California Polytechnic State University, San Luis Obispo (Cal Poly), a master's-granting, teaching-centered institution. There was a more than 60 percent response rate to the survey. The survey results provided insight into the science researchers' data curation awareness, behaviors, and attitudes, as well as what needs they exhibited for services and education regarding maintenance and management of data. It is important that professional librarians understand what researchers both inside and outside their own institutions know so that they can collaborate with their university colleagues to examine data curation needs.

Data curation has been defined as "the active and ongoing management of data through its life cycle of interest and usefulness to scholarship, science, and education … [including] activities [that] enable data discovery and retrieval, maintain its quality, add value, and provide for reuse over time, and this new field includes authentication, archiving, management, preservation, retrieval, and representation."[1] There is a growing demand by taxpayers, government funding agencies, and researchers for open access to data sets. Increased access to research data will allow for verification and replication of results, provide a foundation for additional research, and increase the overall transparency of science. Data curation needs will only become more acute as granting agencies such as the National Science Foundation (NSF) require researchers to deposit underlying data sets along with their published research.[2] There are a number of initiatives currently focusing on the development of infrastructure for and management of massive data sets such as the NSF-funded DataNet Initiative.

Jeanine Marie Scaramozzino is the College of Science & Mathematics, School of Education, and Data & GIS Librarian and Marisa L. Ramírez is the Digital Repository Librarian in the Robert E. Kennedy Library and Karen J. McGaughey is an Assistant Professor of Statistics California Polytechnic State University; e-mail: jscaramo@calpoly.edu, mramir14@calpoly.edu, kmcgaugh@calpoly.edu. *© 2012 Jeanine Marie Scaramozzino, Marisa L. Ramírez, and Karen J. McGaughey, Attribution-NonCommercial (http://creativecommons.org/licenses/by-nc-sa/3.0/) CC BY-NC*

Major scientific journals including *Science* and *Cell* are also developing policies addressing the submission of data sets. These policies include allowing, recommending, or requiring data sets as supplements to completed manuscripts; requiring the data sets to be freely accessible to colleagues; suggesting that data sets be placed in public depositories; and requiring links, accession numbers, and other identifiers that provide clues to the location of data. Interestingly, 71 percent of large publishers (publishers that produce more than 50 journals) and 57 percent of small publishers (publishers that produce less than 50 journals) allow authors to submit underlying data with their publication.[3] It is interesting to note that these publishers, large and small, produce 94 percent of all for-profit and open access journals. However, though publishers may allow for the submission of the data, there are almost no guidelines or details concerning formatting or other data curation issues such as licensing.

The careful stewardship of the underlying research data used in publications is critical, particularly when considering projects in interdisciplinary domains such as environmental science and climate change. Cross-disciplinary endeavors are dependent upon access, discovery, and interoperability of data sets drawn from a variety of sources. However, past studies indicate that most scholars do not have the knowledge required to manage their data effectively.[4] Macdonald and Martinez-Uribe (2010) cite two recently published reports that illustrate this disconnect: Oxford's *Scoping Digital Repository Services for Research Data Management* (2009) and RIN's *Patterns of Information Use and Exchange: Case Studies of Researchers in the Life Sciences* (2009).[5] These studies outline the gaps in researchers' and scholars' knowledge of data curation issues. Currently, researchers who want to submit and share their data lack guidance and training.

Given the lack of data curation awareness in most disciplines, academic libraries have a remarkable opportunity to apply traditional strengths toward collecting and organizing digital research content. According to Choudhury, data curation practices for libraries include viewing "data as collections; data as services; librarians as data scientists; and data centers as the new library stacks."[6] It is therefore crucial for libraries to better understand how science researchers collect, record, and disseminate knowledge and to understand more clearly the library's role in managing data assets effectively. There is a significant relationship between scientific study and scholarly communication. Examination of data management issues will enable a deeper understanding of how libraries can meet researcher needs and how librarians might develop relationships with other data resource providers to facilitate richer, more robust services.[7] This is particularly important given the increasing competition for research funding within the Science, Technology, Engineering, and Mathematics (STEM) fields.[8]

A better understanding of researcher needs and the library's role in data management will not only increase the production of data but will also address patron needs associated with access to data. Patrons' increased needs for digital data assets will influence the selection of library resources and services, resulting in the transformation of librarians into data scientists and libraries into data centers.[9] These data centers may deliver a variety of services, including data curation education, short-term storage, long-term storage, active partnerships with scientists during data creation, and the creation of local, national, and international consortia data networks.

Many academic libraries have infrastructure such as institutional repositories in place to support the acquisition and delivery of locally created digital content. These repositories are the foundational infrastructure that libraries can build upon to serve data needs. Because "librarians can put researchers in touch

with standards applicable to their needs, create a plan for managing the lifecycle of the data in compliance with their grants, create organizing strategies for documentation, files, backups and more," libraries are uniquely poised to provide support and education on the proper curation of scientific data sets.[10] Scientists should not be left to manage digital data on their own; instead, "librarians will have to step forward to define, categorize, and archive the voluminous and detailed streams of data generated in experiments."[11] Many large research universities such as Purdue, Johns Hopkins University, and the University of California at San Diego are investigating institutional approaches to data curation, including the exploration of the role, infrastructure, and services the library should provide for massive data sets generated by researchers. However, data management practices within teaching-centered institutions have not been extensively explored.

As a member of the CSU system, California Polytechnic State University, San Luis Obispo (Cal Poly) has historically been viewed as a teaching institution. Within the past 30 years, there has been a gradual shift from a teaching model to a "teacher-scholar model," where faculty are not only required to teach but are also required to conduct research as part of their retention and promotion process. Cal Poly recently stated, "faculty scholarship, research and creative activity are essential components of the CSU's teaching-centered mission."[12] Given this new teacher-scholar model and the increased focus on campus research productivity, a study of Cal Poly's College of Science and Mathematics (COSAM) faculty was undertaken to determine scientists' current data management activities, assess scientists' level of awareness of data curation issues, identify gaps in scientists' understanding of best practices for maintenance and management of data, and identify education or service opportunities that could enhance and support scientists' data management practices.

Background

Data is the essential raw material of science and a valuable asset on an institutional, disciplinary, and national scale with tremendous potential for integration and reuse.[13] Scientific data sets are often categorized into two groups: data from "Big Science" and data resulting from "Little Science."[14] Big Science describes large-scale research efforts characterized by massive budgets, expensive machines, extensive laboratories, and large numbers of collaborators.[15] Little Science contains some elements of Big Science, but in comparison to Big Science, Little Science operates on "shoestring budgets by unknown pioneers."[16]

Little Science and Big Science enjoy a mutual symbiotic relationship in which both benefit from the activities of the other, making it critical to study the data curation needs of small-scale as well as large-scale research projects. Little Science (i.e. Small Science) stands to benefit most from a concerted data curation effort since Small Science research communities tend to be heterogeneous in the methods and data types applied, without uniform or widely applied data standards, and are not supported by disciplinary repositories.[17] In fact, Small Science is predicted to generate two to three times more data than Big Science in upcoming years, creating a pressing and heretofore unrecognized need for the advancement of data curation best practices.[18] Libraries and librarians now have the unprecedented opportunity to provide the necessary stewardship in the data curation process.

Scientists and scholars are increasingly generating vast amounts of digital content in the form of learning materials, publications, and research data, yet data in digital form is extremely fragile due to limited standards for and adoption of good practices.[19] Most academic libraries support the delivery and maintenance of text-based collections in a variety of print and digital formats, as well as the management and delivery of images, multimedia files, sound, maps, and various other artifacts of

research and culture.[20] University libraries are increasingly recognizing that patrons have as-yet unmet needs for the management of research data sets. Libraries are being called upon to provide value-added services to meet the needs of academic user groups and their corresponding data communities. These value-added services include engaging with scientists during research production cycles; supporting data handling and management; facilitating data deposition; data literacy training and support; collaborating with various offices like campus IT and the grants development office; applying the theory and tools of library and information science to maximize the usefulness of research data; offering services for collection development; representing and linking supporting data management and scholarly communication needs at the beginning of the research process; and facilitating data organization, preservation, and reuse.[21]

Some libraries have taken steps to develop consultation and referral services and to provide technological support systems for publishing data. The same libraries also have taken steps to advocate for responsible and open access to data, while cultivating campuswide partnerships to ensure data stewardship.[22] Many institutions have begun to create positions for digital data librarians and subject data librarians in such areas as chemistry, natural sciences, and GIS. Others now require that subject specialists be versed in data curation, perform campus needs assessments as part of their regular duties, and support the education of their fellow librarians about data curation.[23] Cultural and financial barriers must be removed to support a new sustainable distribution of labor and tasks between data authors, digital curators, data managers, and data users.[24] The key challenges facing many research libraries are both tangible and social in nature: lack of money and resources, lack of faculty interest, lack of shared campus values, and the unwillingness of library staff to be retrained to manage data.[25]

Technologies such as cloud computing, augmented and virtual reality, discovery tools, open source software, and new social networking tools affect nearly all library operations.[26] These new technologies expand the capacity and ability to "collect" data sets, compelling libraries to find new ways to support advances in research and various educational services.[27] It is crucial that libraries seize this valuable opportunity to become recognized as data curation resources in campus communities.

Approach and Motivation

A number of research universities, including MIT and Cornell, have programs in which their libraries play significant roles in Big Science data creation and maintenance processes. Meanwhile, librarians at teaching-centered universities like Cal Poly need to gain further insight into scientists' attitudes and activities in relation to smaller scale data creation and management. As the data output from the Small Science researchers grows, a greater understanding of researcher needs will better inform the approach and nature of data services offered to faculty by the library.

Previous research on faculty data management activities range from data audits used to capture a snapshot of campus technology solutions and digital assets, needs assessments designed to better understand the scope of training, management and data preservation concerns, and data case studies that use interviews to develop a depth of understanding of data creation practices in specific disciplines or fields.[28-30] While these studies provide insight into digital data management and activities, there are some gaps to fill. Data audits provide a snapshot of technological assets available on a campus; but, given the speed of technological change, audits have limits in their ability to direct data services and are specific to the campus at which the audit was conducted. Data curation needs assessments, typically targeting faculty and researchers, provide insight into researcher data

curation behaviors, but the studies often have low response rates, thus making it difficult to draw broad conclusions. Data case studies like the Purdue Data Profiles provide a deep understanding of disciplinary data practices based on detailed interviews with individual faculty. However, given the individual and descriptive nature of case studies, some findings may be specific to the discipline or university, thus limiting the utility of the results and their applicability outside a very specific domain. The Purdue Data Profiles Researchers themselves have stated that "this is not a statistical study (the sample size is neither large nor randomized) or a comprehensive needs assessment but it is a 'deep dive' that allows for valuable insights and establishes an anchor point for more generalized research in the future."[31]

This study focuses on research performed by teacher-scholars and intends to provide insight into current Small Science data curation practices at Cal Poly. Faculty may believe they are informed about data curation, but in practice they may not be using optimal methods to reuse and preserve their data. This statistical, comparative survey of faculty data curation perceptions and behavior will inform libraries of current faculty activities and identify data curation knowledge gaps and strengths.

Methods

The Cal Poly faculty survey was designed to address three major areas of interest within data curation: (a) data preservation; (b) data sharing; and (c) educational needs. These are explained below.

Data Preservation

Data preservation encompasses all of the activities/behaviors that faculty use to preserve both active and past research data. The authors identified six main components of the act of data preservation: (1) the existence of a responsible data management party; (2) data backup; (3) funding for data preservation within grants; (4) data migration to new technologies; (5) data reuse by the individual researcher; and (6) the existence of data preservation plans.

Data Sharing

Data sharing was defined as the act of sharing data with other researchers and was assessed using components (7) and (8): creation of metadata, and data reuse by others.

Educational Needs

Of interest here is what faculty believe are their needs for data curation education. This was assessed with component (9): education on data curation best practices. In addition, educational needs were identified based on faculty weaknesses discovered in the survey.

Survey

The survey was conducted between April 2 and April 22, 2010 at Cal Poly. Cal Poly is a nationally ranked, four-year public institution with just over 19,000 students (approximately 95% undergraduate and 5% post-baccalaureate/graduate) and 1,235 faculty (including part-time faculty).[32] It is one of the 23 campuses in the CSU system and emphasizes comprehensive undergraduate education.

After IRB approval, survey invitations were e-mailed to all 331 College of Science and Mathematics (COSAM) faculty.[33] At Cal Poly, COSAM consists of the departments of Biology, Chemistry, Kinesiology, Mathematics, Physics, Statistics, Liberal Studies, and the School of Education. The e-mail invitation, sent on April 2, 2010, included a brief description of the goals of the survey, a link to the online survey, and information regarding a gift card incentive for completing the survey. An e-mail reminder was sent out one week after the survey was launched. The survey received the full support of the COSAM dean and seven department chairs. The dean also sent out a personal message supporting this research following the initial survey invitation. Similar messages were sent by each of the department chairs. The level

of interdepartmental cooperation may have contributed to the high response rate.

Of the 331 faculty to whom survey invitations were sent, responses were filtered to include only science faculty from the Biology, Chemistry, Kinesiology, Mathematics, Physics, and Statistics departments who engaged in data collection in the course of their research. Our analysis focused exclusively on the 131 tenured or tenure-track faculty (assistant, associate, and full professor status, thus filtering out research assistants, lecturers, and emeritus faculty).[34] As part of the Cal Poly teacher-scholar model, tenure-track and tenured faculty at all levels are expected to engage in professional development programs that demonstrate external validation (such as publishing in peer-reviewed journals and/or obtaining grant funding). These individuals would be the most likely to participate in active research programs and would have the greatest need for data curation education, making their attitudes most relevant. The resulting sample included 82 respondents out of 131 eligible faculty, for a 62.6 percent response rate. Table 1 shows response rates by academic rank. Table 2 shows response rates for each academic department.

TABLE 1
Response Rates by Academic Rank

Rank	Number Responding	Totals in COSAM	Response Rate
ASSISTANT	32	53	60%
ASSOCIATE	26	38	68%
FULL	24	49	49%

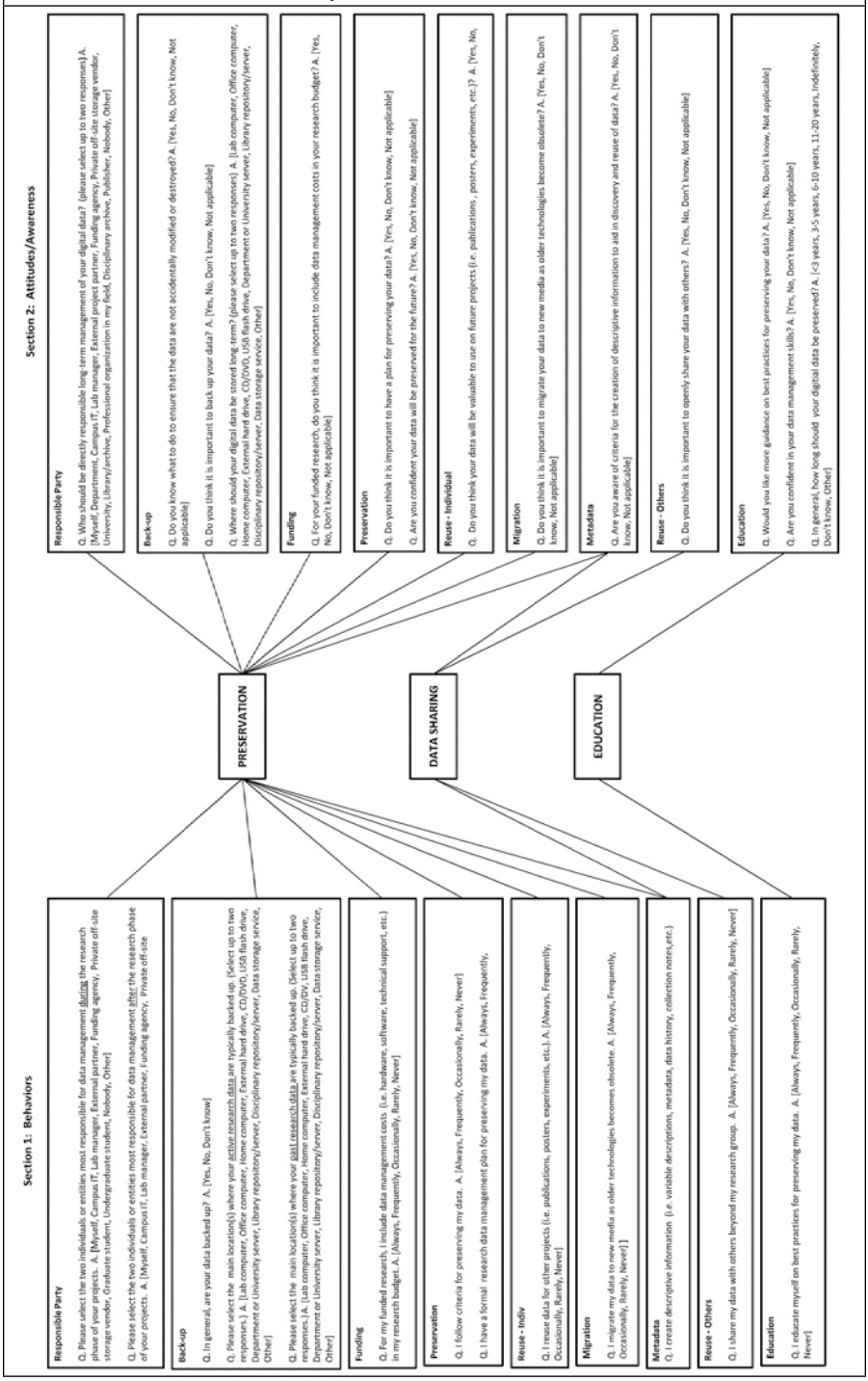
The survey was composed of eighteen questions that were developed to collect information on current and past data management practices/behaviors, as well as opinions/attitudes regarding data management best practices. Specifically, a survey map was constructed to pair faculty behavior and attitude questions addressing the three areas of interest mentioned previously: (a) data preservation; (b) data sharing; and (c) educational needs. Question pairing was done to enable the evaluation of inconsistencies between what faculty members believe is important and what they are actually doing with their data (see figure 1).

The frequency of current and past behaviors for components (3)–(9) were measured using a 5-point Likert scale (Always, Frequently, Occasionally, Rarely, Never). Attitude questions for components (3)–(9) were measured using a dichotomous response (Yes/No) with an option to select "Don't know" or "Not applicable" if the respondent had no knowledge regarding the question or believed the question was not applicable to his/her data collection experiences. Questions for behavior and attitude for components (1) and (2) allowed respondents to "select the top two" from a given list of answer choices, since answers to these questions could be dependent upon the specific research project in which the researcher was engaged. The survey format, question wording, length of the survey, use of an incentive, and use of an online survey tool, Survey Monkey, were all considerations in the construction of the survey. The researchers aimed to reduce the

TABLE 2
Response Rates by Academic Department

Department	Number Responding	Total in Each Department	Response Rate
BIOLOGY	15	30	50%
KINESIOLOGY	8	11	73%
CHEM/BIOCHEM	15	26	58%
PHYSICS	17	29	59%
MATHEMATICS	17	34	50%
STATISTICS	10	14	71%

FIGURE 1
Survey Questions Were Employed to Determine Faculty Behaviors and Attitudes as They Relate to Data Curation Activities



burden on respondents in an effort to increase the response rate and eliminate bias. The survey had built-in skip logic that made sure that respondents only saw relevant questions.

The survey was pretested on a representative group of nine Cal Poly COSAM science faculty and department chairs. Changes were made to the survey format and question wording was updated to reflect concerns and eliminate points of confusion as indicated by the pretesters. While appropriate measures were taken to reduce any potential sources of bias, with a response rate of 62.6 percent there is the possibility of bias due to nonresponse. The individuals who did not respond to the survey may have answered differently from those who did respond to the survey. Additional sources of bias may have been introduced by allowing individuals to skip questions, scroll backward and forward, change their answers, and exit at any time.

Results

This section is divided into three subsections: (a) data preservation; (b) data sharing; and (c) educational needs.

Data Preservation

Component (1) addresses the existence of a responsible data management party for the preservation of both current (active) research and past research data. Respondents were asked to select up to two entities that are directly responsible for the management of both active and past research data. Figure 2 shows the results for both active and past research data. For active research, 93 percent of respondents report that they are personally responsible for the management of their data, with another 40 percent indicating that an undergraduate student may be responsible. For past research data, 97 percent report that they are personally responsible for data management, while 40 percent indicate that no one was responsible. When asked who should be directly responsible for the long-term management of their data, 95 percent of respondents believe they should be personally responsible (see figure 3). Interestingly, faculty generally do not believe that entities such as libraries, campus IT, external project partners, professional organizations, and disciplinary archives should be responsible for the long-term

FIGURE 2
Individuals or Entities Most Responsible for Management of Active and Past Research Data (Respondents Selected up to Two Entities)

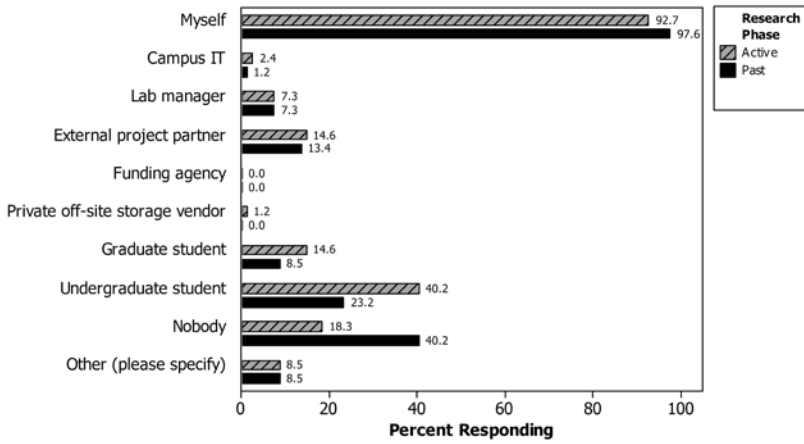
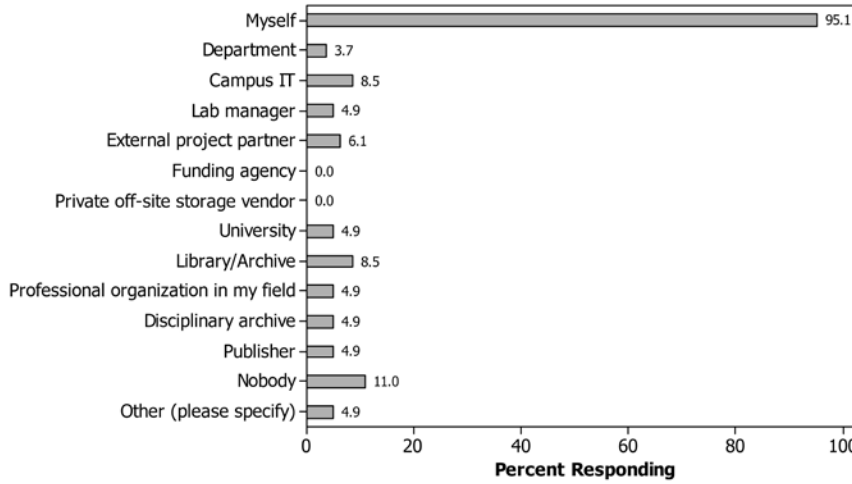


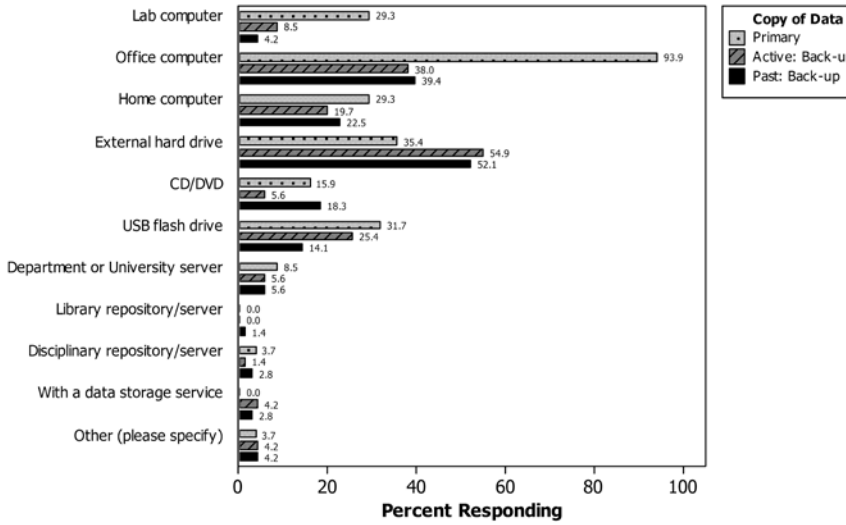
FIGURE 3
Faculty Attitudes Regarding who Should be Responsible for Long-term Management of Data (Respondents Selected up to Two Entities)



management of their data, though these agencies may be better equipped to manage long-term digital storage.

Component (2) addresses the issue of data backup. Are researchers regularly backing up their active and past research data? If so, what are the primary locations for data backup? The survey questions asked respondents to select up to two locations for the storage of the primary

FIGURE 4
Storage Locations for the Primary Copy of Research Data, the Backup Copy of Active Research Data, and the Backup Copy of Past Research Data



data files and up to two locations for the backup copies. As a point of reference, 94 percent of respondents report storing the primary copy of their research data on their office computer, while 30 percent to 35 percent also report storing the primary copy on a lab computer, home computer, USB flash drive, or external hard drive. A total of 86 percent of respondents report their data are indeed backed up. Figure 4 shows the locations for primary data storage along with data backup locations for active and past research data. The number one location for the backup copy is an external hard drive, holding 55 percent of active and 52 percent of past research data, respectively. The second most popular location for the backup copy is an office computer, with 38 percent and 39 percent, respectively. A majority (58%) of researchers claim to be backing up their active research at least weekly (17% any time there are changes, 14% daily, and 27% weekly), while for past research data, almost half (48%) report backing up either quarterly, annually, or any time there are changes (24% every time there are changes, 14% annually, and 20% quarterly).

When asked about data backup, 99 percent of the respondents believe it is important to back up their data, but only 60 percent know what to do to make sure their data are not accidentally modified or destroyed. When asked to select the top two locations for long-term data storage, almost half of respondents (48%) report data should be stored long-term on an external hard drive, while 37 percent believe the storage space should be an office computer, and 28 percent see a department or university server as the best storage space for data.

Component (3) addresses funding behaviors and attitudes toward data curation. Respondents were asked if they believe it is important to include data management costs in their research budgets and if they actually budget for these costs. Only 34 percent of faculty members believe it is important to include data management costs in their grant applications, and only 20 percent of respondents report always, frequently

FIGURE 5
Distribution of Faculty Who Always, Frequently, Occasionally, Rarely, or Never Have Preservation Plans in Place Based on Their Belief in the Importance of Such Plans

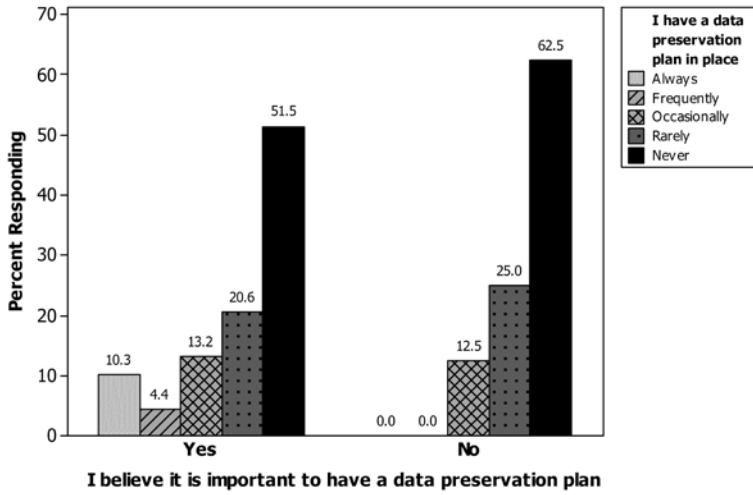
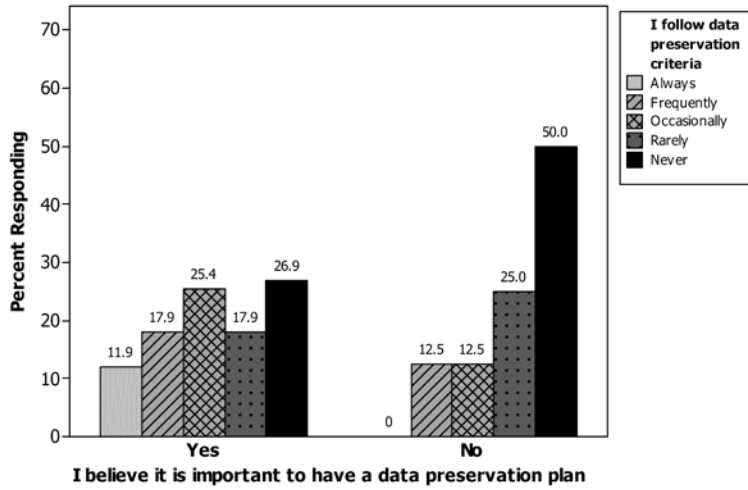


FIGURE 6
Distribution of Faculty Who Always, Frequently, Occasionally, Rarely, or Never Follow Data Preservation Criteria Based on Their Belief in the Importance of Preservation Plans



or even occasionally factoring in these costs.

To address component (4), faculty were asked if they believe it is important to migrate their data to new media as older technologies become obsolete. An example of hardware migration includes transferring data stored on an obsolete computer hard drive to a flash drive or external hard drive. Fully 89 percent believe it is important to save older data on newer mediums. When asked if they actually do transfer old data files to new media technologies, 79 percent responded with always, frequently, or occasionally.

Reuse of data by the individual researcher is the focus of component (5). Faculty respondents were asked whether they believe their data will be valuable to them for future research projects and whether they have reused their data for other projects. Over 90 percent of participants reported occasionally reusing their data, and 80 percent believe their data will be valuable for future projects. Component (6) deals with the use of formal preservation plans. A total of 84 percent of respondents believe it is important to have a data preservation plan in place. Of this group, fewer than 15 percent report always or frequently having such a plan (see figure 5). In addition, only 30 percent who believe it is important to have a data preservation plan report always or frequently following best practices in data preservation (see figure 6). Fewer than half of researchers (40%) are confident that their data will be preserved for the future.

Data Sharing

Component (7) addresses the creation of metadata to facilitate data reuse and sharing. Only 20 percent of faculty report being aware of criteria for the creation of descriptive information to aid in discovery and reuse of data. Less than 10 percent report being both knowledgeable on and always or frequently using said criteria.

For component (8), faculty respondents were asked about their beliefs in the importance of sharing their data with others and the frequency with which they share their research data. Over 65 percent of respondents believe it is important that

they openly share their data and that their colleagues do the same. Of those who believe it is important, fewer than half (48%) report always or frequently sharing data with those outside their research group.

Educational Needs

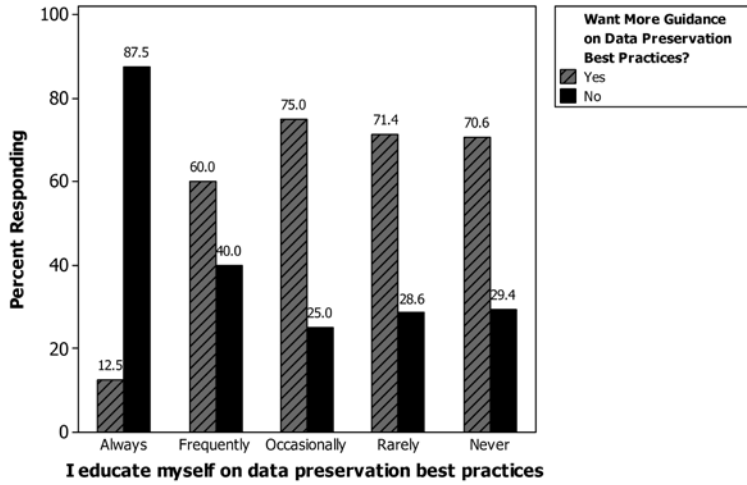
In the last component (9), educational needs and wants are addressed. Participants were asked how confident they were in their data management skills, if they educate themselves on data preservation best practices, and if they would like more guidance on data preservation best practices. Fully half (50%) of respondents report they are either not sure or not confident in their data management skills. Of these responders, only 7 percent report they always or frequently educate themselves on best practices for preserving data. Among the sample as a whole, only 20 percent report they always or frequently educate themselves on data curation best practices. For those individuals who occasionally, rarely, or never educate themselves, we see a strong desire by more than 70 percent of responders for more guidance and education on best practices (see figure 7).

Discussion

Anecdotal information would lead one to believe that faculty members conducting Small Science do not back up their data or have comprehensive data management plans. However, according to the results of our study, many Small Science researchers regularly back up their data. Nonetheless, faculty lack proper backup procedures, and data management is not generally an accepted component of their workflow. The majority of faculty who responded to the survey recognize that they need guidance to improve their data management activities. In addition, these faculty members are open to additional education on data management. However, the library is not perceived as a resource to provide this service.

Faculty members see themselves as primarily responsible for their research data during and after data collection. During the research phase, however, they share this responsibility with their students. Student involvement is a result of the emphasis on undergraduate research at Cal Poly. Given that students are responsible for the management of data during the collection phase, students are

FIGURE 7
Distribution of Responses for More Guidance on Data Preservation Best Practices Based on Level of Current Data Preservation Self-Education



an important consideration during the development of data curation services and library outreach programs.

Our results indicate that most faculty store the primary copy of research data on an office computer or external hard drive, and they back up active and past inactive data in these same two places. With the backup copy in the same location as the primary copy, failure of the hardware could mean a devastating loss of data. In addition, more than a third of respondents use a USB drive, lab computer, or home computer to store the primary copy of their data. These locations may not be administered by a trained computer technician or online backup service, which means that research data is at risk. In these cases, a single computer failure or lost USB drive could lead to a catastrophic loss. Faculty members do believe that they should store their data with their department or university server, though they do not always do so. While two-thirds of respondents believe that sharing data is important, storing data on a closed server prevents such sharing. Additionally, trust in the department or university server for long-term storage may be misplaced, particularly if no formal agreements or practices are in place to curate data over time.[35] As mentioned above, few respondents reported the library as a potential source for long-term storage services. This is surprising, particularly as the Cal Poly Robert E. Kennedy Library does have an institutional repository infrastructure in place to house data.

Respondents report valuing data management plans when it comes to managing their academic research. Of course, there are costs associated with the care and management of data. However, according to the results of the survey, faculty rarely account for these costs within grant applications. This is an important educational avenue that the library or other campus entities, such as a Grants and Development Office, could explore.

The majority of respondents reported reusing their data, which is consistent with the widespread belief that research data has intrinsic value. Ironically, while the majority of researchers believe that colleagues should share their data, only a minority of respondents actually share their own data with individuals who did not help in gathering the data.

Based on the results of this study, respondents appear to need additional guidance for creating metadata, preserving and sharing data, writing data management plans and gaining an improved understanding of data curation best practices. While the majority of respondents feel confident that their data will be preserved for the future, their responses demonstrate that only a minority are following best practices and are educated on data preservation issues. Faculty recognize the need to be more informed about data management practices, and they are open to educational opportunities to increase their knowledge on the subject. However, the library is not perceived as a locus for assistance in the data curation life cycle. Instead, faculty see themselves as the responsible parties for maintaining their data. The challenge for libraries is to determine the data curation services that can assist faculty the most, while also creating opportunities to promote library strengths and expertise. This would then demonstrate the primary role libraries could play in managing researchers' data. These services must be of value to faculty and be viable from a financial and a human resources standpoint.

Educational initiatives developed by the library would inform faculty on data curation issues. Based on the results of the study, faculty indicated interest in gaining access to data curation educational materials. To meet this need at Cal Poly, the Kennedy Library now hosts a data curation research guide featuring practical recommendations based on sound practice.[36] The guide includes information on the basics of data management, educational resources, backup practices, ethical/legal/copyright issues, funder requirements, the creation of data manage-

ment plans, links to data repositories and databases, and links to other data management resources. The research guide has been promoted during presentations to new faculty and graduate students and used in relevant seminars organized by Cal Poly's Office of Research and Graduate Programs. Feedback has been positive, and the online usage statistics indicate growing interest in the resource.

While this first step is modest, the ultimate hope of the researchers is to develop additional services to broaden faculty awareness of data curation issues that span a wide array of disciplines. Because data issues are applicable to researchers at any stage of their career, there are opportunities to educate both tenured and untenured faculty, as well as graduate and undergraduate students. If faculty are engaging in good data curation practices, their students who assist with data collection will also benefit as future scholars.

A number of ancillary benefits were also derived from the distribution of this survey. Informal word of mouth generated interest among faculty who want to learn more about data curation issues. Consequently, the library was asked to present information about library services, resources, and infrastructure that support research and grant writing to faculty and graduate students. Additionally, faculty are contacting librarians for help with data management plans (requesting lists of discipline-specific repositories, to deposit data in the library's institutional repository, and data management plan templates and writing assistance).

Additional attention was generated from a broad cross-section of groups across campus. The campus Grants Development Office regularly handles numerous Department of Defense and Office of Naval Research grants that require data management plans. As a result of this survey, we have discovered a keen interest in coordinating workshops on data management plans with the Grants Development Office, the Center for Teaching and Learning, and the Research and Graduate Programs Office.

A recent university reorganization has created promising future developments for data curation at Cal Poly. The Kennedy Library and the campus Information Technology Services division merged and now report to the University Vice Provost for Information Services. This development will lead to collaborative opportunities that will shape data curation services for research faculty. At Cal Poly, most of the librarian job descriptions now include data curation activities. This is important, particularly because these liaisons will now be charged with a more active role in curating faculty research. In addition, a new Library Data and GIS Services Program is in development to support the data needs on campus including the creation of the "Data Studio."[37]

Conclusion and Recommendations

Whether produced by Small Science or Big Science, all research data is scientific capital. As it becomes common scientific practice to deposit these assets in data repositories, it is important for librarians to understand scientists' data activities to better support them.[38] By conducting a survey on university teacher-scholars and their data curation behaviors and attitudes, we discovered that, while Small Science faculty report following some data management practices, they do not necessarily adhere to the best practices. Faculty members recognize the need to become better informed on data management issues and are open to increased educational opportunities on this topic. However, they do not perceive libraries as a source of data management expertise or as the best place to store academic research data. Nonetheless, the library is a fountain of knowledge whose potential has not yet been fully tapped. Data curation is an avenue to demonstrate how integral the library can be in the research process.

As libraries and librarians understand the opportunities afforded by integrating data librarianship into their services and recognize the value they can provide,

they will need to hone skills, forge new partnerships with scientists and data managers, and become a vital part of the scholarly record.[39] We suggest that library leaders take the time to consider and answer the following questions: What types of data curation educational opportunities can librarians take advantage of? What types of librarians should be receiving this education? How do libraries successfully become part of the research dynamic? In what ways can librarian-researcher partnerships be fostered?

As the demand for research data sets continues to increase, tools supporting preservation, discovery, access, and education will need to evolve along with the raw results of research. If libraries wish to play a role in this quickly changing arena, they will need to foster a culture of flexibility, immediacy, and service. The way forward is inevitably through a mix of cross-institutional and cross-disciplinary structures that can take multiple forms. These can fall within the categories of both Big and Small Science and can range from national and international organizations, to smaller regional centers that may be colocated within existing centers, to specialized collections housed on individual campuses that serve a well-defined community.[40] Needs will be best identified and matched with capabilities by fostering librarian-researcher partnerships and establishing programs for mutual engagement and education. We suggest that librarians and researchers work together to identify potential solutions to data management challenges, consolidate assets, and collectively advocate for campus adoption of data management policies and support.

Coevolution between librarians and researchers will allow libraries a greater ability to influence concomitant transformations in science creation and scientific sharing workflows, scholarly communication models, and support infrastructure. Thus, we envision a reciprocal flow of influence: librarians influencing the data practices of scientists and data practices of scientists influencing the services provided by libraries.

Acknowledgements

We would like to thank Michael Miller, Vice Provost for Information Services and CIO; Anna Gold, University Librarian; Tim Strawn, Director of Information Resources and Archives; Dr. Phil Bailey, College of Science and Mathematics Dean; and the COSAM Department Chairs, faculty, and staff for their invaluable support of this research. We would also like to thank Anna Gold, Melissa Cragin, and Susannah Kopecky for their constructive feedback on our manuscript.

Notes

1. The definition of "data curation" posted on the Master of Science: Specialization in Data Curation University of Illinois Graduate School of Library & Information Science (GSLIS) Web site, available online at www.lis.illinois.edu/academics/programs/ms/data_curation [accessed 1 October 2010].

2. National Science Foundation, "Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans," *National Science Foundation (NSF) News* (May 2010), available online at www.nsf.gov/news/news_summ.jsp?cntn_id=116928&org=NSF&from=news [accessed 1 October 2010].

3. PARSE.Insight Consortium, *Science Data Infrastructure Roadmap* (June 2010), available online at www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf [accessed 1 October 2010].

4. Lesley Freiman, Catharine Ward, Sarah Jones, Laura Molloy, and Kellie Snow, *Incremental Scoping Study and Implementation Plan: A Pilot Project for Supporting Research Data Management* (2010), available online at www.lib.cam.ac.uk/preservation/incremental/documents/Incremental_Scoping_Report_170910.pdf [accessed 1 October 2010].

5. Stuart Macdonald and Luis Martinez-Uribe, "Collaboration to Data Curation: Harnessing Institutional Expertise," *New Review of Academic Librarianship* 16 (2010): 4–16, available online at www.tandfonline.com/doi/abs/10.1080/13614533.2010.505823 [accessed 1 November 2010]; Luis

Martinez-Uribe, *Findings of the Scoping Study and Research Data Management Workshop* (2008), available online at http://ora.ouls.ox.ac.uk/objects/uuid:4e2b7e64-d941–4237-a17f-659fe8a12eb5 [accessed 1 November 2010]; RIN Disciplinary Case Studies in the Life Science Project, *Patterns of Information Use and Exchange: Case Studies of Researchers in the Life Sciences* (Nov. 2009): 1–56, available online at www.rin.ac.uk/our-work/using-and-accessing-information-resources/patterns-information-use-and-exchange-case-studie [accessed 1 November 2010].

6. Sayeed Choudhury, "Rethinking Scholarly Communication: Building Data Curation Infrastructure" (presentation at the meeting of Utah Campus Infrastructure Day, Salt Lake City, Utah, Mar. 13, 2009), available online at www.it.utah.edu/leadership/research/ciday/2009/notes/choudhury.pdf [accessed 1 October 2010].

7. Melissa Cragin, "Shared Scientific Data Collections: Use and Functions for Scientific Production and Scholarly Communication" (doctoral dissertation, University of Illinois at Urbana Champaign, 2009).

8. Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, *Data Audit Framework Methodology* (May 2009), available online at www.data-audit.eu/DAF_Methodology.pdf [accessed 1 November 2009].

9. Choudhury, "Rethinking Scholarly Communication," 7.

10. Tracy Gabridge, "The Last Mile: Liaison Roles in Curating Science and Engineering Research Data," *Research Library Issues: A Bimonthly Report from ARL, CNI, and SPARC* 261:15–21, available online at www.arl.org/bm~doc/rli-265-gabridge.pdf [accessed 1 November 2009].

11. Scott Carlson, "Lost in a Sea of Science Data," *The Chronicle of Higher Education* v. 52, no. 42, available online http://chronicle.com/article/Lost-in-a-Sea-of-Science-Data/9136 [accessed 25 May 2012].

12. W. David Conn and Bruno Giberti, *Our Polytechnic Identity in the 21st Century: WASC Capacity and Preparatory Review Report* (Dec. 2009), available online at http://wasc.calpoly.edu/pdfs/cpr/cpr_essays_web.pdf [accessed 12 January 2010].

13. Melissa Cragin, *Trends and Opportunities: Data Curation and Data Literacy* (presentation at the Robert E. Kennedy Library, California Polytechnic State University, June 2009).

14. "Big Science" was originally coined by Alvin M. Weinberg in his 1961 work "Impact of Large-Scale Science on the United States," *Science* 134 (3473): 161–64.

15. Derek J. de Solla Price, *Little Science, Big Science* (New York: Columbia University Press, 1963), 161; Weinberg, "Impact of Large-Scale Science on the United States."

16. de Solla Price, *Little Science, Big Science*, 3.

17. "Little Science" is now referred to as "Small Science"; Melissa Cragin, Carole Palmer, Jacob Carlson, and Michael Witt, "Data Sharing, Small Science and Institutional Repositories," *Philosophical Transactions of the Royal Society A* 368 (2010): 4023-4038.

18. Carlson, "Lost in a Sea of Science Data."

19. Cragin et al., "Data Sharing, Small Science and Institutional Repositories."

20. Anna Gold, "Cyberinfrastructure, Data, and Libraries, Part 2 Libraries and the Data Challenge: Roles and Actions for Libraries," *D-Lib* 13 (2007), available online at www.dlib.org/dlib/september07/gold/09gold-pt2.html [accessed 20 January 2009].

21. National Science Foundation, *Long-lived Digital Data Collections Enabling Research and Education in the 21st Century* (NSB-05-40) (Dec. 2005), available online at www.nsf.gov/pubs/2005/nsb0540/ [accessed 20 January 2009].

22. Gold, "Cyberinfrastructure, Data, and Libraries."

23. Examples available at http://www.lisjobs.com/jobseekers/results.asp?search=data+librarian&submit2=Search&anyallexact=exact. [Accessed 25 May 2012].

24. Carlson, "Lost in a Sea of Science Data"; National Science Foundation, *Long-lived Digital Data Collections.*

25. Catherine Soehner, Catherine Steeves, and Jennifer Ward, *E-Science and Data Support Services: A Study of ARL Member Institutions* (Aug. 2010), available online at www.arl.org/bm~doc/escience_report2010.pdf [accessed 15 January 2011].

26. ACRL Research Planning and Review Committee, "2010 Top Ten Trends in Academic Libraries: A Review of the Current Literature," *College & Research Libraries News* 71 (2010): 286–92, available online at http://crln.acrl.org/content/71/6/286.short [accessed 15 January 2011].

27. Carole Palmer, Lauren Teffeau, and Carrie Pirmann, *Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development* (Jan. 2009), available online at www.oclc.org/research/publications/library/2009/2009-02.pdf [accessed 20 March 2009].

28. Philip Lord and Alison Macdonald, *Data Curation for e-Science in the UK: An Audit to Establish Requirements for Future Curation and Provision* (2003), available online at www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf [accessed 20 March 2009]; Cuna Ekmekcioglu and Robin Rice, *Edinburgh Data Audit Implementation Project Final Report* (Jan. 2009), available

online at http://ie-repository.jisc.ac.uk/283/ [accessed 15 January 2011]; Neil Jerrome and Jacobs Breeze, *Research Data Management Imperial College DAF Pilot Final Report* (Mar. 2009), available online at http://ie-repository.jisc.ac.uk/307/ [accessed 15 January 2011]; Panaylota Polydoratou, *UCL Data Audit Framework Pilot Implementation Final Report* (Apr. 2009), available online at http://ie-repository.jisc.ac.uk/404/ [accessed 15 January 2011]; Cynthia Gering, Lisa Schmidt, Nancy Fleck, Catherine Foley, Traci Gulick, Shawn Nicholson, Cindy Straus, and Don Ries, *Final Report: MSU Digital Curation Planning Team* (Jul. 2010), available online at http://msudcp.archives.msu.edu/wp-content/uploads/2010/07/DCP-Final-Report1.pdf [accessed 15 November 2010].

29. Margaret Henty, Belinda Weaver, Stephanie Bradbury, and Simon Porter, *Investigating Data Management Practices in Australian Universities* (Jul. 2008), available online at www.apsr.edu.au/investigating_data_management [accessed 15 November 2009]; Kenji Takeda, Mark Brown, Simon Coles, Les Carr, Jeremy Frey, Peter Hancock, and Graeme Earl, "Institutional Data Management Blueprint" (presentation at the meeting of the University of Southampton, U.K., Mar. 29, 2010), available online at www.southamptondata.org/uploads/7/3/0/0/730051/idmboxford29march2010.pdf [accessed 1 October 2010]; Freiman, et al., *Incremental Scoping Study and Implementation Plan*.

30. Neil Beagrie, Robert Beagrie, and Ian Rowlands, "Research Data Preservation and Access: The Views of Researchers," *Ariadne* 60 (Jul. 2009), available online at www.ariadne.ac.uk/issue60/beagrie-et-al/ [accessed 1 October 2010]; Andre Holzner, Peter Igo-Kemenes, and Salvatore Mele, "Data Preservation, Reuse and (Open) Access: A Case Study in High-Energy Physics" (presentation at the PARSE.Insight Workshop, Darmstadt, Germany, June 21, 2009), available online at www.parse-insight.eu/downloads/PARSEInsight_event200909_casestudy_HEP.pdf [accessed 1 October 2009]; Stuart Madnick, MacKenzie Smith, and Kate Clopeck, *How Much Information? Case Studies on Scientific Research at MIT* (Jun. 2009), available online at http://hmi.ucsd.edu/pdf/HMI_Case_Summary.pdf [accessed 1 January 2010]; Key Perspectives Ltd., *Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long-term Viability* (Jan. 2010), available online at www.dcc.ac.uk/scarp [accessed 1 March 2010]; RIN Disciplinary Case Studies, *Patterns of Information Use and Exchange*; Beagrie, Beagrie, and Rowlands, "Research Data Preservation and Access."

31. Michael Witt, "Eliciting Faculty Requirements for Research Data Repositories" (presentation at the meeting of the 4th International Conference on Open Repositories, Georgia Tech, Atlanta, Ga., June 4, 2009), available online at http://hdl.handle.net/1853/28509 [accessed 1 October 2009].

32. Numbers located on page 72 of the Office of Institutional Planning and Analysis's *Cal Poly Fall 2009 Fact Book*, available online at www.calpoly.edu/~ipa/publications_reports/factbook/fbfall09.pdf [accessed 25 May 2012].

33. Number located on the *2009–2010 College of Science and Mathematics Roster (Update 4/6/10)*. Retrieved from the California Polytechnic State University, San Luis Obispo, College of Science and Mathematics Web site: http://cosam.calpoly.edu/ [accessed 1 August 2009]. This roster is no longer available online.

34. Numbers provided by a personal communication with the College of Science and Mathematics personnel analyst, 28 May 2010.

35. It must also be noted that each COSAM department is served by two separate ITS entities: departmental and campus. Some faculty may not have a server available to them and/or ITS policies and staff may constantly be changing.

36. Launched October 2010, the College of Science and Mathematics Data Management research guide is in development and can be viewed online at http://libguides.calpoly.edu/data [accessed 25 May 2012].

37. The library "learning commons" environment brings together library, technology, and other services to foster informal, collaborative work, and social interaction. The vision for our "data commons," the Data Studio, is to provide a similar atmosphere focused on data and metaliteracy. Additional information on the Data Studio can be found at http://libguides.calpoly.edu/datastudio. Additional information on metaliteracy: Thomas Mackey and Trudi Jacobson, "Reframing Information Literacy as a Metaliteracy," College and Research Libraries 72 (2011): 62-78. Available online at http://crl.acrl.org/content/72/1/62.full.pdf+html [accessed 25 May 2012].

38. Christine Borgman, Jillian C. Wallis, and Noel Enyedy, "Building Digital Libraries for Scientific Data: An Exploratory Study of Data Practices in Habitat Ecology" (paper presented at the meeting of the 10th European Conference on Digital Libraries, Alicante, Spain, Sept. 17–22, 2006, available online at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.95.230 [accessed 1 October 2009].

39. Gold, "Cyberinfrastructure, Data, and Libraries."

40. Academic Research Libraries, *To Stand the Test of Time Long-term Stewardship of Digital Data Sets in Science and Engineering*, Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe (2006), available online at www.arl.org/pp/access/nsfworkshop.shtml [accessed 1 October 2009].