Research Data Management
At the University of Colorado Boulder

Recommendations in support of fostering 21st Century research excellence

November 15th, 2012

Prepared by the Vice Chancellor for Research's Data Management Task Force

Patricia Rankin (chair), *Associate Vice Chancellor for Research*

Barbara Buttenfield, *Professor, Department of Geography*
Ruth Duerr, *Data Stewardship Program Manager, National Snow & Ice Data Center*
Thomas Hauser, *Director, Research Computing*
Andrew Johnson, *Assistant Professor and Metadata Librarian, University Libraries*
Jack Maness, *Assistant Professor and Director of Sciences Department, University Libraries*
Mark Parsons, *Senior Associate Scientist, National Snow & Ice Data Center*
Hari Rajaram, *Professor, Department of Civil, Environmental, and Architectural Engineering*
Rich Shoemaker, *Director, Nuclear Magnetic Resonance Spectroscopy Facility*
Kimberly Stacey, *Research Data Manager, Research Computing*
Alex Viggio, *Faculty Information Systems Lead Developer, Office of Faculty Affairs*
Jina Choi Wakimoto, *Associate Professor and Director of Metadata Services, University Libraries*

University of Colorado **Boulder**

Table of Contents

I. Executive Summary

Research stakeholders—including funding agencies, universities, and researchers—have identified research data management (RDM) as key to the efficiency and value of the research process and to enabling research innovation.

To remain competitive in research excellence, as emphasized in the *Flagship 2030* strategic plan, CU-Boulder must foster robust work practices for RDM and invest in the infrastructure, both social and technical, needed to successfully support RDM at the campus level.

The Data Management Task Force (DMTF) assessed the current state of RDM at CU-Boulder, using surveys, review of peer institutions, and synthesis of existing data lifecycle models. Based on their findings, the DMTF recommends that CU-Boulder take the following steps to develop the RDM infrastructure needed to compete and lead on a global scale:

- **Endorse and establish an ethical, open data policy** at the campus level as a first and fundamental step toward successful RDM.

- **Highlight RDM** on campus by encouraging Deans and Chairs to value and acknowledge RDM-related activities in the promotion and tenure process.

- **Develop** outreach and communication efforts around RDM.

- **Develop clear policies and procedures for research data** that address issues of ownership, access and preservation, ethical and legal concerns (e.g., privacy), and roles and responsibilities.

- **Encourage** faculty to consider various forms of Open Access publishing, a necessary compliment to open data sharing, as noted by the National Science Board (2011).

- **Create a Research Data Services (RDS) unit** to begin immediately providing basic RDM services to researchers including referrals to existing resources, data management planning, storage for active data, and archiving of completed data sets.

- **Invest in appropriate personnel and technology** and establish a sustainable funding model for both of these integral pieces.

- **Leverage** existing solutions for data management including systems and local, national, or international data repositories, where they exist.

II. Introduction and Background

The University of Colorado Boulder's *Flagship 2030* strategic plan emphasizes "fostering research excellence" and "investing in the tools for success" as initiatives central to a national comprehensive research university's ability to compete and lead on a global scale (University of Colorado, 2008). In a growing number of disciplines, research excellence depends on computationally intensive methods, networked and distributed environments, instruments capable of acquiring data in enormous quantities, and reanalysis of existing data (Hey, Tansley, and Tolle, 2009). As a result of these developments, researchers now produce a vast amount of digital data, and in turn rely more than ever on data to enable new forms of inquiry and discovery (Office of Science and Technology Policy, 2012).

For research data to be this springboard for continued discovery, the rapidly growing body of research data must be curated, archived, and preserved to ensure discoverability, access, and reuse over the short and long term (Committee on Science, 2009). Furthermore, broader access to well-managed research data increases the verifiability and reproducibility of findings, while reducing duplicate data collection efforts (Schofield et al., 2009). Potential benefits of data sharing for individual researchers include increased citation rates for their publications when they share their data sets (Piwowar, Day, and Fridsma, 2007). The full potential of these benefits depends largely on the degree of open data sharing (Stodden, 2009) and links to publicly available scholarly publications. Therefore, data management and sharing should be considered in discussions about open access to peer-reviewed literature (National Science Board, 2011).

Many U.S. funding agencies now recognize the benefits of improving research data sharing and access, and have enacted policies and recommendations to promote proper data management (see Appendix A). In January of 2011, the National Science Foundation (NSF) released one of the most prominent of these policies, including a requirement for all grant proposals to provide a data management plan (National Science Foundation, 2011). As funding agencies like the NSF continue to develop and strengthen policies concerning research data, research universities must in turn provide support for data management to ensure their researchers remain competitive for grants, and to encourage innovation and discovery through new forms of research and scholarship (Macdonald & Martinez-Uribe, 2010). Many institutions have developed models for delivering data management services and resources at the campus or university level (see Section IV). As their experiences show, support for research data management at this level requires the collaboration of a number of stakeholders including libraries, computing and IT units, contract and grant offices, existing data centers and repositories, and researchers themselves.

Given these challenges, successful data management services would require substantial investment in **sociotechnical infrastructure** (Star and Ruhleder, 1996; Bowker et al., 2010). In the context of data management, the sociotechnical infrastructure encompasses not only the software platforms, the access policies, and the data itself, but also the knowledge discovered, exchanged, and produced as a consequence of the technical services and capabilities which support effective data management. An effective sociotechnical infrastructure can be said to "fit the needs, activities, and contexts of the people who use it, as well as those of the people who create it, operate it, and contribute to its content." (Van House et al., 2003). Another increasingly common term is **cyberinfrastructure** as defined by Edwards et al. (2007, p. 6), but we prefer the term sociotechnical infrastructure for its explicit recognition of the social element.

To address the evolving landscape of research data management, the Data Management Task Force (DMTF) was created to identify research data management needs at CU-Boulder and provide recommendations for how best to develop services and infrastructure to meet those needs. The DMTF first met in late July 2011 and continued to meet monthly until October 2012. The group, chartered by Stein Sture, Vice Chancellor for Research, was initially small, but was expanded to represent campus research more broadly. The final members include Babs Buttenfield, Ruth Duerr, Thomas Hauser, Andrew Johnson, Jack Maness, Mark Parsons, Harihar Rajaram, Patricia Rankin, Rich Shoemaker, Kimberly Stacey, Alex Viggio, and Jina Choi Wakimoto. See Appendix B for the full DMTF charter.

The DMTF engaged in a number of information-gathering efforts described later in this report. The DMTF also investigated the ISO standard, *Open Archives Information System Reference Model*, and the following seven data lifecycle models. These models provided a conceptual framework for data management infrastructure and services recommendations.

- Data Curation Centre (DCC) Curation Lifecycle Model (Higgins, 2008)
- Data Documentation Initiative (DDI) Combined Life Cycle Model (Structural Reform Group, 2004)
- Australian National Data Service (ANDS) Data Sharing Verbs (Burton & Treloar, 2009)
- DataONE Data Lifecycle (Michener & Jones, 2012)
- UK Data Archive Data Lifecycle (UK Data Archive, 2012)
- Research360 Institutional Research Lifecycle (Jones, 2011)
- Capability Maturity Model for Scientific Data Management (Crowston & Qin, 2011)

See Appendix C for the full comparison and associated data management needs that arose from this investigation.

For the purposes of this report, "data" is defined broadly as the digital representation of information generated at any stage of the research process in a formalized manner suitable for communication, interpretation, or processing generated at any stage of the research process (adapted from the *Reference Model for an Open Archival Information System*, http://public.ccsds.org/publications/archive/650x0b1.pdf). Data can be produced from a variety of processes (e.g., observation, experimentation, simulation, derivation, compilation), represented in numerous forms (e.g., text, numerical, multimedia, model, software, discipline-specific, instrument-specific), and stored in many digital formats (e.g., ASCII, PDF, SPSS, Excel, TIFF, Java, FITS, CIF, ZVI) (MIT Libraries, 2009). The scope of this definition includes data from disciplines in the sciences, social sciences, and humanities.

III. State of the Campus

The task force began by assessing the current landscape for research data management at CU-Boulder. In addition to surveying researchers about their data management practices and needs, the task force reviewed a report from the Libraries' participation in the 2011-2012 Association of Research Libraries (ARL) E-Science Institute, examined existing resources like the CU-Boulder institutional repository, and identified efforts related to research data management already underway in the CU-Boulder Libraries and Research Computing.

1. Survey of Campus Researchers' Data Management Practices

The Data Management Task Force (DMTF) members gathered information about current researchers' data management practices via an online campus-wide survey. The survey was sent to the campus community in January 2012 via the Faculty and Research e-Memo list consisting of 4,411 names, the Buff Bulletin, and targeted e-mails to individuals. There were 148 complete responses. While this is a small response rate, it could nonetheless reflect the relatively small number of people currently on campus who are sufficiently versed in data management practices to be able to respond. The twenty-two question survey was designed to communicate about the DMTF to campus researchers, raise awareness about important questions related to data management, gather information about the data and data management practices across the institution, and offer an opportunity for respondents to give feedback about services and priorities.

The survey responses help describe the current state of data management activities across local research disciplines. The various survey questions gave information about:

- Respondents and their research areas
- Types of research data generated
- Current storage amounts and projected growth
- Length of time research data will need to be accessible
- Data and metadata formats for storage
- Maintenance of data or metadata documentation
- Implementation of formal data management plans
- People managing data
- Storage and backup technology being used
- Proportion of data that is sensitive, confidential, or proprietary
- Interest in different types of data management services

The survey was written with broad language to ensure that respondents from all domains could reply. Given the diverse nature of respondents' roles, departments, and research areas, the survey responses can be applied statistically to a large majority of the university's population of researchers. All schools and colleges at CU-Boulder were represented in the responses. More specifically, the scope of the responses includes 38 departments, 6 institutes, 5 centers and 2 programs representing 337 research areas. Responses also came from Accounting and Business Support, Athletics, CU Museum, Law Library, Museum of Natural History, Office of Information Technology, University Information Systems and University Libraries.

The results of the survey indicate that the majority of respondents do not have a clear understanding of how best to manage their data in an easy and effective manner.

The following is a summary of findings from the survey (see Appendix D for the full report):

a. Across campus, **respondents need assistance with data management**. An overwhelming percentage of respondents stated that they lack data management plans and/or metadata, and they would like help with planning activities and assistance creating and maintaining data management plans and metadata. The need for data management consultation exists in nearly every academic discipline across the campus.

b. An individual CU Boulder researcher will use **many different file types and data types for research data**. No single respondent listed less than four different file and/or data types in their responses to the survey. Nearly 150 unique file types were listed as relevant and important across all responses, and in all cases respondents use at least two different types of data such as documents, spreadsheets, digital media, etc. This diversity in data type presents significant challenges for campus-wide RDM.

c. The **total data stored by researchers at CU Boulder varies** regardless of research area or role. Data volume is not characterized by discipline; there was no statistical probability of seeing higher or lower amounts of data for particular departments and institutes. In any given area of campus people are likely to have more than 1MB and less than 10TB of data.

d. A majority of **respondents collected the total of their stored data over a period of three years**.

e. On average, **researchers** at CU Boulder **need access to their data for at least five years**.

f. Most respondents, regardless of research area or department, **do not maintain metadata or data documentation**. There was no statistical difference between departments in their responses to these questions. Only 36% of respondents stated that they do have metadata and these numbers are consistent for all departments across the campus.

g. **Few researchers have a data management plan** and this is true across all reported research areas and departments. There was no statistical difference between departments in their responses to these questions. Only 24% of respondents stated that they do have data management plans and these numbers are consistent for all departments across the campus.

   The vast majority of respondents at CU Boulder **manage their own research data**. An overwhelming number of respondents stated that they themselves are responsible for management of their data. A considerably smaller number of respondents stated that others within their department or others outside their department or the university have a hand in managing their data.

2. University Libraries' Report from the 2011-2012 ARL E-Science Institute

The task force reviewed a report resulting from the CU-Boulder Libraries' participation in the 2011-2012 Association of Research Libraries (ARL) E-Science Institute. The E-Science Institute was designed in 2011 to help research libraries develop a strategic agenda for e-research support with a particular focus on the sciences. CU-Boulder was part of the first cohort participating in this institute.

The report, titled *The Strategic Initiative for Research Data Support and Services at the University of Colorado Boulder Libraries,* concludes that in order to address funding agency mandates, University leadership must build support by communicating to campus constituents, supporting involvement of cross-functional experts, clarifying expectations, and funding the development of tools to support data sharing and preservation as appropriate.

The report notes that it is in the best interest of the University to add value to the research data produced by CU-Boulder through effective data management and preservation. To

address data management issues and coordinate efforts, the report recommends establishing a Research Data Services unit at CU-Boulder. This unit would represent the cross-functional collaboration necessary to solve complex data management, sharing, and preservation needs. The report outlines a potential partnership between the Office of Information Technology (OIT), the Office of the Vice Chancellor for Research, Research Computing (RC), and the University Libraries. The report also discusses potential University Libraries service roles in data management and curation, metadata and ontologies, and outreach. See Appendix E for the full report.

3. Digital Collections of Colorado at CU-Boulder (Institutional Repository)

The task force also examined the possibility of using the existing CU-Boulder institutional repository to house research data. The libraries of the University of Colorado and Colorado State University systems have partnered to develop a digital repository called the Digital Collections of Colorado (http://digitool.library.colostate.edu) in order to provide public access to the research outputs of each campus. Each library is managing its own individual institutional repository (IR), but the various IRs are connected through shared infrastructure including the DigiTool (ExLibris Ltd.) repository software platform. The CU-Boulder institutional repository (http://ucblibraries.colorado.edu/repository) accepts submissions from CU faculty members, departments, other campus units, and faculty-sponsored students. Acceptable content includes pre- and post-prints of scholarly articles, presentations, white papers, reports, theses, dissertations, dissertation supplements, and other forms of scholarship.

While most IRs (and IR software platforms) are designed primarily to house and provide access to documents and document-like digital objects, the task force determined that some data sets could be considered appropriate content for inclusion in the CU-Boulder IR. The IR is not capable of serving as a full-lifecycle data management system, but it could be used in some cases to store and provide public access to final, completed data sets.

The inclusion criteria for data sets in the IR would likely include the following:

a. **Data that can be made publicly available**. The CU Boulder Digital Collections is intended to be an “open access” repository, and the DigiTool software does not allow for the fine-grained access controls needed to house sensitive data (e.g., personally identifiable information from human subjects).

b. **Data sets that are in a fixed form** and not intended to be updated or revised. The DigiTool software does not have versioning capabilities.

c. **Data that are not intended to be analyzed, manipulated, or queried** within the DigiTool interface. Data files would need to be downloaded and imported into external software for any analysis, manipulation, or searching within data sets.

d. **Data that are well-documented.** Documentation for data sets should go far beyond the simple descriptive metadata required for ingest into the DigiTool software (e.g., Title, Author, Date, etc.). In many cases, separate documentation files will need to be included in order to allow potential users to understand and reuse the data.

If the CU-Boulder IR is to be used to provide access to some data sets, the above criteria should be explicitly stated in policies and guidelines. Additional policies concerning format migration and any curation activities that could allow data sets to be understood and used beyond a few years should also be addressed. The following are examples of similar policies and guidelines from other institutions:

"Guidelines for Research Dataset Contributions in DSpace@MIT"
http://libraries.mit.edu/dspace-mit/build/policies/dataset-guidelines.html

"eCommons@Cornell Data Deposit Policy"
http://ecommons.library.cornell.edu/policy.html#data

4. Research Computing and University Libraries Current Research Data Management Efforts

Finally, the task force identified efforts that Research Computing and the Libraries have undertaken in recent years by utilizing existing resources and personnel. Members of the Libraries and Research Computing created an ad-hoc group to explore RDM issues on campus in response to a growing awareness of the importance of RDM as well as events like the release of the NSF Data Management Plan requirements. This group developed a website (http://data.colorado.edu) with RDM resources as well as basic data management plan consulting services. Both units recognize the need to further develop these services as soon possible, but that more comprehensive RDM services would require further resources not currently available.

IV. Review of Peer Institutions

The task force chose to review six peer institutions and assess their models of research data management services, including organizational, funding, personnel, policy, and technical considerations. These models have helped inform the task force's recommendations on a number of levels.

1. Organizational and Service Models

All peer institutions utilize collaborative organization and service models for the management of research data services, and all include institutional equivalents of the Libraries, Office of Information Technology, and the Associate Vice Chancellor for Research. One peer additionally includes the Office of Contracts and Grants, and several leverage the work of graduate students in programs of Library and Information Science and campus research institutes that have been active in data management for many years.

- **Cornell University** has a strong support and infrastructure called The Research Data Management Service Group (RDMSG, https://confluence.cornell.edu/display/rdmsgweb/Home). The RDMSG is jointly sponsored by the Senior Vice Provost for Research and the University Librarian, has a faculty advisory board (nine faculty members from various disciplines and two ex officio from Office of Research Integrity and Assurance and Office of Sponsored Programs) and a management council (seven members—two librarians, two discipline faculty, two from computing/IT, one from an institute). It serves as a clearing-house to point researchers to campus services that range from intellectual property and metadata to storage security and high performance computing.

- **Purdue University** takes a collaborative approach: the Libraries provide consulting and metadata support; campus Information Technology provides storage and research computing support. The Executive Committee includes the Dean of Libraries, the VP of Research, and the VP of Information Technology.

- The **University of Illinois at Urbana-Champaign** includes the Office of the CIO, IT, OCG, Libraries, and library school students. Libraries provide data management plan support and a "Scholarly Commons" space, which offers advanced software and specialized hardware for data analysis and other activities.

- The **University of North Carolina-Chapel Hill** has a Data Management Committee of the University Library and the Health Sciences Library, in

partnership with a variety of University offices and groups, including its library school, campus IT, many research institutes, and the Carolina Digital Repository. It also provides a new data repository that is a joint project between two research centers, one in life sciences, the other in information science.

- The **University of Virginia Library** has a Scientific Data Consulting Group (SciDaC) that includes three full-time equivalent staff, works closely with subject librarians, the university attorney, the institutional repository team, the Chief Information Officer, and the Office of the VP for Research. SciDaC provides consultation in data management plans, metadata, and offers workshops and events to faculty and graduate students.

- The **University of Wisconsin-Madison** provides a Research Data Services unit that is a collaboration among the Libraries, Department of IT, the CIO office, the Graduate School, and the School of Library and Information Studies to assist researchers with data curation needs. Researcher support is provided by the Digital Curation Team, made up of twelve members (six librarians, three academic technologist/IT, and graduate students in the School of Library and Information Studies).

2. Funding Models

Institutions in the review demonstrated a variety of funding models. Some units, primarily libraries and campus IT, absorb the cost, while others receive funding or obtained new personnel from campus administration. Reallocation of existing personnel is also a common theme.

- **Cornell University** provides a great many data services in a very distributed manner. Some are fee-based to researchers; others are provided by institutes and centers, which presumably absorb the cost.

- **Purdue University** absorbed initial project costs in addition to receiving grants from the Institute for Museum and Library Services. Subsequent proposals to the campus resulted in additional funding for new full- and part-time positions in the Libraries and campus IT:
    - Digital Library Software Developers
    - Digital Data Repository Specialist
    - Metadata Specialist (20%)
    - Digital Archivist (25%)

- Graduate Assistant (50%)
- HUB Liaison (25%)
- Technical Project Manager (10%)
- Software Developer (50%)
- Middleware Developer (25%)
- Security Expert (10%).

These central services provide 100GB of free storage and support for the life of grant-funded projects. Additional storage can be purchased for about $2100 per TB per year. Assistance is provided to write storage costs into grant proposals.

- At the **University of Illinois at Urbana-Champaign** the Provost provides funding for speakers on data management, recently including an NSF assistant director. The campus has also funded two positions in the libraries, one librarian and a data curation specialist developer. The office of the CIO pays for the storage costs and server maintenance of their institutional repository (which will provide static, "published" data), and the libraries provide the staffing. The library and information science program is also heavily involved.

- The **University of North Carolina Chapel Hill** offers at least four different options for data management, two of which absorb costs, one of which utilizes cost-recovery, with the fourth obtaining its funding through an NSF grant, with plans to move into a cost-recovery model in the future. The ITS-Research Computing solution provides three years of storage for $620 per TB after the first 100GB.

- The **University of Virginia** Libraries and campus IT share the cost. The Libraries provide software developers and direct user services, and IT provides the technical infrastructure and system administrators.

- The **University of Wisconsin at Madison** provides many storage options with a variety of cost-recovery models (http://researchdata.wisc.edu/manage-your-data/data-backup-and-integrity/). Service personnel costs are absorbed by units, including a digital curation consultant, subject librarians, and a research services librarian in the libraries, and IT consultants in departments, funded by campus IT.

See Appendix F for a full list of institutions with available research data management websites.

V. Recommendations

In line with CU-Boulder's strategic plan and developments surrounding research data management at the national and international levels (as outlined in Section II of this report), the DMTF makes the following recommendations. They are based on analyses of existing data lifecycle models, the current state of the campus with respect to RDM, and the efforts underway at peer institutions.

In general, the DMTF recommends that CU-Boulder highlight and encourage RDM, officially create and support a virtual organization to begin providing research data management services, develop governance and procedures for that organization, and establish a sustainable funding model for the infrastructure and personnel required for full-lifecycle RDM.

Specifically, the DMTF recommends that CU-Boulder:

1. Highlight and encourage research data management

The academic leadership of CU-Boulder, (the Office of the Vice Chancellor for Research (OVCR) being a likely leader), should adopt and promote a set of principles following the model set forth by the National Science Board (NSB) but expanded to include all disciplines represented on this campus. The NSB recommends the following principles (National Science Board, 2011):

"1. Openness and transparency are critical to continued scientific and engineering progress and to building public trust in the nation's scientific enterprise. This applies to all materials necessary for verification, replication and interpretation of results and claims, associated with scientific and engineering research.

2. Open Data sharing is closely linked to Open Access publishing and they should be considered in concert.

3. The nation's science and engineering research enterprise consists of a broad array of stakeholders, all of which should participate in the development and adoption of policies and guidelines.

4. It is recognized that standards and norms vary considerably across scientific and engineering fields and such variation needs to be accommodated in the development and implementation of policies.

5. Policies and guidelines are needed for open data sharing which in turn requires active data management.

6. All data and data management policies must include clear identification of roles, responsibilities and resourcing.

7. The rights and responsibilities of investigators are recognized. Investigators should have the opportunity to analyze their data and publish their results within a reasonable time."

The DMTF recommends the OVCR work in concert with the Provost and the Associate Vice Chancellor for Academic Affairs as well as Deans and the Boulder Faculty Assembly (BFA) in order to **promote these principles by encouraging**:

- **Alignment of faculty review systems** with the NSB guiding principles. Deans, Chairs, faculty governance, and tenure and promotion evaluation committees should consider the intellectual value of data creation, sharing, and stewardship in their evaluative work (e.g., faculty who demonstrate the positive impact of their shared data or open publications are rewarded at review cycles);

- **Faculty to consider various forms of Open Access publishing** and archiving of publications, and to consider policies like those of Harvard and other Tier One universities given the close connection between Open Access publishing and Open Data sharing noted by the NSB in point 2 above, and;

- **Faculty to adopt policies for sharing data in the most open way possible,** given the norms of each discipline and the rights and responsibilities of individual researchers. The DMTF recognizes that the results of some research cannot be shared openly, but encourages faculty to do so whenever possible.

2. Formally create a Research Data Services Organization

The DMTF recommends that Research Computing and University Libraries continue their work in establishing Research Data Services (RDS). The Provost, OVCR, Dean of Libraries and CIO should formally recognize and support this new virtual organization's efforts. With current resources, RDS should support a minimum suite of basic services as described below in Section 2.b Significant additional funding is needed to implement all of the fundamental data services that the task force believes the University should

provide (see Section 4). As part of its charge, RDS will investigate, select, and implement appropriate software, tools, and processes as resources are identified.

RDS will consult and involve campus and CU partners as appropriate. Regular partners could include the Technology Transfer Office for intellectual property issues, IT Security for data security needs, the Office of Contracts and Grants for matters concerning grant submissions, and the Institutional Review Board for privacy concerns. A unit like the Office of Contracts and Grants, involved in the initial stages of proposal submission, could create procedures that would allow for the notification of RDS for data management plan consulting. This initial consulting would assist researchers in acquiring grant funding by helping them comply with data management plan requirements. Additional benefits of such an interaction include:

- **Early interaction with researchers** to ensure the accuracy of the information in the data management plan,

- Benefit the long-term process of managing data by **ensuring best practices** are being employed during the planning stage, and

- **Build awareness among researchers** about the campus services and support professionals in RDS.

RDS will also collaborate with and leverage the existing data management expertise of CU-Boulder affiliated research institutes and data centers (IBS, LASP, NSIDC, etc.). Other possible collaborations could include working with the Office of Faculty Affairs to explore the potential of linking open data sets to researcher profiles in the CU-Boulder VIVO instance.

The following is an outline of the recommended structure of and services provided by RDS:

a. Organization of Research Data Services

The Research Data Services organization should consist of three groups:

i. Research Data Services Operations – This group would provide actual services to researchers and would include representatives from Research Computing, University Libraries, and other groups with relevant expertise, perhaps most notably the National Snow and Ice Data Center (NSIDC). Members would be added as services expand, requested positions are approved, and personnel are hired. This group would report to the Research Data Executive

Committee described below. Initial leadership of this group should consist of co-coordinators from University Libraries and Research Computing.

ii. Research Data Executive Committee – The Research Data Executive Committee (RDEC) would advise, support, and authorize the Research Data Services Operations group's efforts to develop the necessary services for research data management at CU-Boulder. The RDEC would meet regularly to discuss budget and personnel needs and make decisions that cross organizational boundaries. Members could include the Director of Research Computing, the Associate Vice Chancellor for Research, and an Associate Dean or other senior personnel from University Libraries.

iii. Research Data Advisory Committee - The Research Data Advisory Committee (RDAC) would direct research data governance, procedures and policies at the campus level, while ensuring broad representation of faculty members and other stakeholders including experts to best offer solutions.

b. Basic services RDS will provide with current resources

The services listed below will be provided as a common good to the campus community, using currently available resources. These services would only be provided for data that are open and free of Health Information Portability and Accountability Act (HIPPA), Family Educational Rights and Privacy Act (FERPA), and other legal restrictions. It should be noted that these are minimal services. Should demand increase and as technology needs inevitably change, additional resources and personnel will be required to sustain even this level of operation.

i. Data Management Planning, Consulting, and Training
RDS will review existing data management plans and provide general feedback about basic tools, practices, resources, and services. RDS will also assist in planning an estimated budget for the data management requirements of projects, and will provide informal education and training workshops for researchers on data management best practices, as well as associated issues.

ii. Research Data Storage
RDS will provide storage services to CU Boulder's researchers. These storage services are currently under development. An up-to-date list of the storage services is provided at https://www.rc.colorado.edu/resources/storage.

iii. Metadata and Documentation Consulting
RDS will refer researchers to existing standards, resources, and/or personnel.

iv. Support for Sharing Data Among Collaborators During Active Projects
RDS will provide referrals to existing resources (e.g., third parties such as Globus Online).

v. Periodic Value Assessment
RDS will consult with researchers on the selecting/deselecting of data for archiving and/or continued storage.

vi. Archiving
RDS will provide referrals to existing repositories/archives. For some small static datasets, RDS will provide persistent storage and access via the Digital Collections of Colorado (DCC) institutional repository. It should be noted that the DCC does not have all the features a true data archive system employs. For example, data could be at risk because there are no integrity checks.

vii. Data Visualization/Analysis Support
RDS will refer researchers to existing tools that are compatible with existing infrastructure, and/or existing personnel.

c. Necessary services RDS could provide with additional resources

Long term preservation and access is essential to good research in all disciplines and is increasingly seen as a responsibility of researchers and of a modern research institution, as evidenced by increasing data management requirements from funding agencies. To meet the minimal standards of these emerging requirements, RDS should be able to provide the following services to ensure data preservation, access, and reuse. These services would require additional resources as described later in Section IV of this report.

i. Ingest
RDS should provide a basic service for taking in data from researchers. This service should include checks to ensure the integrity of digital data files and the quality and completeness of metadata and documentation needed to ensure long-term data preservation, access, and reuse. This service should also involve an agreement between the researcher and RDS as to what exactly will be preserved, for how long, and what this preservation will entail.

ii. Full Archiving
While existing repositories and archives may be appropriate for small static data sets, certain types of data, and data from particular disciplinary communities,

RDS should provide a full archival storage and retrieval solution for all remaining CUB-produced data. Such an archive should comply with the Open Archival Information System (OAIS) Reference Model as laid out in the ISO 16363:2012 standard for trustworthy digital repositories (International Standards Organization, 2012). By definition, such an archive would require constant monitoring, planning, and maintenance in order to ensure reliable, long-term preservation of and access to data.

iii. Access

RDS should be able to accept data with a variety of access needs that may or may not change throughout the data lifecycle. Data for active projects should be securely sharable among collaborators across the campus, the university, and those outside the campus. RDS should provide appropriate access controls for data with ethical and legal concerns (e.g., privacy) in compliance with regulations like HIPPA and FERPA. Whenever possible, RDS should provide and promote reliable widespread public access to data through a variety of means including sophisticated user interfaces, search engine optimization, and interoperability with semantic web technologies.

iv. Curation

Curation is necessary for data preservation and archiving. It involves maintaining, preserving, and adding value to research data throughout their lifecycle. Curation enhances the long-term value of existing data by making them available and understandable for further use and reuse. Curation also reduces threats to the long-term value of data and mitigates the risk of digital obsolescence. RDS will need to provide curation services for accepted data if RDS is to remain viable.

v. Citation Support

Data citation is precise reference to the exact data used in a particular study. When done well, it also provides measurable credit to data creators and others. For data citation to be precise and persistent requires that the cited data remain available or that there be description of how data may have changed or have been retired. Data citation is aided by the use of registered persistent identifiers, such as Digital Object Identifiers (DOIs). There is a small cost to obtaining and registering these identifiers, but the greater cost is with ensuring the identifiers remain current and persistent. This is largely a curation function. RDS should be able to support both the registration of persistent identifiers and the curation required to enable long-term citation.

3. **Develop research data governance and procedures**

Lead by OVCR, CIO, Deans and BFA, the Research Data Advisory Committee (see Recommendation 2a above) should determine the individuals and/or units responsible for creating procedures or policies in order to achieve the following:

- **Clearly define "researchers" and "research data"** (i.e., to whom and to what the following policies and procedures will apply).

- **Clarify ownership of intellectual property rights for research data** created by researchers employed at CU-Boulder (i.e., whether the researcher(s), the institution, the funding agency, or the public owns research data).

- **Provide legal** (e.g., HIPAA, ITAR) **and ethical** (e.g., privacy) **guidance** for researchers collecting and/or using data with these considerations.

- **Recommend security measures** for research data, especially data with ethical and legal considerations.

- **Determine appropriate periods of retention** for research data that account for existing disciplinary/community norms and funding agency policies. Provide procedures for archiving and preserving data as well as de-accessioning and disposing of data.

- **Work with broad faculty representation to define what types of data** (e.g., raw data, publishable data, metadata) **should be shared** and at what point in the research process they should be shared (e.g., at completion of study, upon publication of associated articles, after an embargo period). This definition will need to account for the wide variety of data sizes, formats, and disciplinary practices.

- **Promote widespread access to research data** while accounting for disciplinary and community norms, and ethical and legal considerations.

- **Determine the roles and responsibilities of researchers and the institution** in complying with procedures and policies for research data.

- **Define business models** for sustained data curation and preservation.

- **Develop a mechanism that incorporates data management planning into the grant submission process** and provides a mechanism to involve RDS when data management consultation is needed.

4. Establish a sustainable sociotechnical infrastructure necessary for full-lifecycle data management

In order to provide the necessary data management services enumerated in Recommendation 2c—ingest, full archiving, access, curation, and citation support—**fully integrated social and technical infrastructures will be needed**. Review of peer institutions suggest that sustaining a sociotechnical infrastructure necessary to provide research data services could require as many as 5 dedicated FTEs, including the skills of system developers, architects and administrators, metadata and curation specialists, information professionals such as librarians, web developers, and/or other types of liaison personnel.

No technical systems currently exist that would meet all data management needs at the Boulder campus without significant development. A preliminary review of the data management technology at other comparable institutions that generate similar volumes and diversity of data confirms this notion. The DMTF does not anticipate any being available in the next several years. Therefore, either an initial influx of resources dedicated to system development or a formal "outsourcing" arrangement with another institution that is developing their own system would be required. Other scenarios can be imagined, but would require creativity and championing from campus leadership. The Research Data Executive Committee should seek external funding for the development of an institutional archive solution for data sets. It should be anticipated that the development phase may require additional resources, particularly in system programming, but that the need for these resources may diminish as development transitions to maintaining and optimizing an established system.

In order to determine the level of resources required to begin initial development of a local system**, RDS Operations should evaluate several available systems and recommend the appropriate technical and social infrastructure needed** to develop and sustain them. If significant development and system administration are required to build a system for testing, RDS Operations will request resources and personnel to accomplish the goal of evaluating products. These recommendations should be directed at the Research Data Executive Committee.

Finally, the DMTF suggests that the Vice Chancellor for Research work with each college to either create a new position, or assign relevant duties to existing positions, that will serve as a liaison with campus research data management personnel. This staff person should have a good understanding of the specific research data management needs of the cohort of faculty they represent and possess knowledge that will help to facilitate the delivery of the most appropriate services and technology.

VI. Conclusion

The DMTF, as detailed in this report, concludes that the development of full-scale research data management capabilities should be an immediate priority for CU-Boulder. This conclusion is based on information from a number of sources, including national and international reports, funding agency policies and guidelines, scholarship on data management and data lifecycles, input from CU-Boulder researchers and other campus stakeholders, and efforts underway at peer institutions. In particular, the DMTF recommends that CU-Boulder take the following steps toward building a full-scale research data management program:

- **Promote research data management**
- **Formally create a Research Data Services group, reporting structure, and advisory committee**
- **Develop research data governance and procedures**
- **Establish a sustainable funding model for infrastructure and personnel for full-lifecycle data management**

By taking these steps, the DMTF believes CU-Boulder will make significant progress toward providing the tools necessary to continue a tradition of research excellence into the 21st Century.

VII. Appendices

Appendix A. Funding Agency Policies and Guidelines for Data Management

Funding Agency	Policies/Guidelines
Center for Disease Control and Prevention	Releasing and Sharing Data http://www.cdc.gov/od/foia/policies/sharing.htm
Department of Defense	Department of Defense Instruction 3200.14, Principles and Operational Parameters of the DoD Scientific and Technical Information Program http://www.dtic.mil/whs/directives/corres/pdf/320014p.pdf Department of Defense Directive 3200.12, Scientific and Technical Information (STI) Program (STIP) http://www.dtic.mil/dtic/pdf/customer/STINFOdata/DoDD_320012.pdf
Department of Energy	Department of Energy Standard Research Terms and Conditions http://www.nsf.gov/pubs/policydocs/rtc/doe_708.pdf Atmospheric Radiation Measurement (ARM) Climate Research Facility Data Sharing and Distribution Policy http://www.arm.gov/data/docs/policy Developing Data Management Policy and Guidance Documents for your NARSTO Program or Project http://cdiac.ornl.gov/programs/NARSTO/DM_develop_guide.pdf
Department of Health and Human Services Office of Research Integrity	Guidelines for Responsible Data Management in Scientific Research http://ori.hhs.gov/images/ddblock/data.pdf
Department of the Interior	Data Resource Management, Departmental Manual, Series 17, Part 378 Departmental Manual available at: http://elips.doi.gov/elips/

Environmental Protection Agency	Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by the Environmental Protection Agency http://www.cdlib.org/services/uc3/datamanagement/funding.html Survey of EPA and Other Federal Agency Scientific Data Management Policies and Guidance 2010 http://cendievents.infointl.com/SDM_062910/docs/EPA_Policy_and_Guidance_SDM_Report.pdf
Institute of Education Sciences (Department of Education)	Policy Statement on Data Sharing in IES Research Centers http://ies.ed.gov/funding/datasharing_policy.asp
Institute of Museum and Library Services	Specifications for Projects that Develop Digital Products http://www.imls.gov/assets/1/AssetManager/DigitalProducts.pdf
National Aeronautics and Space Administration	NASA Earth Science Data and Information Policy http://science.nasa.gov/earth-science/earth-science-data/data-information-policy/ NASA Earth Science Data Rights and Related Issues http://science.nasa.gov/earth-science/earth-science-data/data-information-policy/data-rights-related-issues/ NASA NASA Guidebook for Proposers Responding to a NASA Research Announcement (NRA) or Cooperative Agreement Notice (CAN) http://www.hq.nasa.gov/office/procurement/nraguidebook/proposer2010.pdf
National Endowment for the Humanities	Data Management Plans for NEH Office of Digital Humanities Proposals and Awards http://www.neh.gov/files/grants/datamanagementplans.pdf General Terms and Conditions for Awards http://www.neh.gov/grants/manage/general-terms-and-conditions-awards-awards-issued-may-2009-or-later#data
National Institute of	National Institute of Food and Agriculture, U.S. Department of

Food and Agriculture (Department of Agriculture)	Agriculture, Terms and Conditions, Small Business Innovation Research Grants Program http://www.nifa.usda.gov/business/awards/sbir2010_05-05-2010_final.pdf
National Institutes of Health	Final NIH Statement on Sharing Research Data http://grants1.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html Data Sharing Regulations/Policy/Guidance Chart for NIH Awards http://grants1.nih.gov/grants/policy/data_sharing/data_sharing_chart.doc NIH Data Sharing Policy and Implementation Guidance http://grants1.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm Frequently Asked Questions on Data Sharing http://grants1.nih.gov/grants/policy/data_sharing/data_sharing_faqs.htm
National Institute of Standards and Technology (Department of Commerce)	National Institute of Standards and Technology Guidelines, Information Quality Standards, and Administration Mechanism http://www.nist.gov/director/quality_standards.cfm
National Oceanic and Atmospheric Administration (Department of Commerce)	"NAO 212-15: Management of Environmental Data and Information" http://www.corporateservices.noaa.gov/ames/administrative_orders/chapter_212/212-15.pdf "NOAA Data Documentation Procedural Directive" https://geo-ide.noaa.gov/wiki/index.php?title=Data_Documentation_PD "NOAA Data Management Planning Procedural Directive" https://geo-ide.noaa.gov/wiki/index.php?title=Data_Management_Planning_PD

	"NOAA Data Sharing Procedural Directive" https://geo-ide.noaa.gov/wiki/index.php?title=Data_Sharing_for_NOAA_Grants_PD
National Science Foundation	Dissemination and Sharing of Research Results http://www.nsf.gov/bfa/dias/policy/dmp.jsp Award & Administration Guide (AAG) Chapter VI.D.4 http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4 Grant Proposal Guide (GPG) Chapter II.C.2.j http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp Data Management & Sharing Frequently Asked Questions (FAQs) http://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp Biological Sciences Directorate (BIO) Directorate-wide Guidance http://www.nsf.gov/bio/pubs/BIODMP061511.pdf Computer & Information Sciences & Engineering (CISE) http://www.nsf.gov/cise/cise_dmp.jsp Education & Human Resources Directorate (EHR) http://www.nsf.gov/bfa/dias/policy/dmpdocs/ehr.pdf Engineering Directorate (ENG) Directorate-wide Guidance http://nsf.gov/eng/general/ENG_DMP_Policy.pdf Geosciences Directorate (GEO) Directorate-wide Guidance http://www.nsf.gov/geo/geo-data-policies/index.jsp Geological Sciences Directorate (GEO) Division of Earth Sciences http://www.nsf.gov/geo/ear/2010EAR_data_policy_9_28_10.pdf

	Geological Sciences Directorate (GEO) Integrated Ocean Drilling Program http://www.nsf.gov/bfa/dias/policy/dmpdocs/geo_iod.pdf Mathematical and Physical Sciences Directorate (MPS) Division of Astronomical Sciences (AST) Advice to PIs on Data Management Plans http://www.nsf.gov/bfa/dias/policy/dmpdocs/ast.pdf Mathematical and Physical Sciences Directorate (MPS) Division of Chemistry (CHE) Advice to PIs on Data Management Plans http://www.nsf.gov/bfa/dias/policy/dmpdocs/che.pdf Mathematical and Physical Sciences Directorate (MPS) Division of Materials Research (DMR) Advice to PIs on Data Management Plans http://www.nsf.gov/bfa/dias/policy/dmpdocs/dmr.pdf Mathematical and Physical Sciences Directorate (MPS) Division of Mathematical Sciences (DMS) Advice to PIs on Data Management Plans http://www.nsf.gov/bfa/dias/policy/dmpdocs/dms.pdf Mathematical and Physical Sciences Directorate (MPS) Division of Physics (PHY) Advice to PIs on Data Management Plans http://www.nsf.gov/bfa/dias/policy/dmpdocs/phy.pdf Social, Behavioral and Economic Sciences Directorate (SBE) Directorate-wide Guidance http://www.nsf.gov/sbe/SBE_DataMgmtPlanPolicy.pdf Social, Behavioral and Economic Sciences Directorate (SBE) Data Archiving Policy for the Division of Social and Economic Sciences (SES) http://www.nsf.gov/sbe/ses/common/archive.jsp
Smithsonian Institution	Smithsonian Environmental Research Data Management Policy http://www.serc.si.edu/research/longterm_data/policy2.aspx Sharing Smithsonian Digital Scientific Research Data from

	Biology http://www.si.edu/content/opanda/docs/Rpts2011/DataSharingFinal110328.pdf
U.S. Geological Survey (Department of the Interior)	USGS National Climate Change and Wildlife Science Center Data Policies and Guidance https://nccwsc.usgs.gov/?q=content/data-policies-and-guidance

Appendix B. Charter for the Data Management Task Force

1. Work to pull together disparate but critical entities and expertise in the CU-Boulder community.
2. Act as a nexus for leading data management efforts.
3. Make recommendations about the storage and curation of digital data produced in the course of CU-Boulder based research.
4. Address the roles of individual researchers, departments and institutes, staff, and the university as a whole.
5. Consider a wide array of data during this process (observational, experimental, clinical, simulation, instrument).
6. Evaluate: (Analysis of survey)
    a. Data sets that vary substantially in terms of size
    b. Appropriate security for different data sets
    c. Which data sets should be retained and for what length of time
    d. Governance issues such as data ownership, stewardship, access and sharing
    e. Necessary policies; and complications that might arise through collaborations with other entities
7. Address storage and maintenance issues in both the short and long-term, and potential funding models for each.
8. Provide specific recommendations about how CU-Boulder investigators can respond to NIH and NSF policies (although its mandate is not restricted to particular funding sources, or limited to funded research).
9. Review policies and practices at other universities as well as the national context in formulating its recommendations for CU-Boulder.

Appendix C. Data Lifecycle Models Comparison and Associated Data Management Needs

(Information used to create this table came from the following report: Alex Ball. 2012. *Review of Data Management Lifecycle Models.* REDm-MED Project Document redm1rep120110ab10. University of Bath.)

Data Lifecycle Models							**Data Management Needs**
DCC	*DDI*	*ANDS*	*DataONE*	*UKDA*	*Research360*	*CMM*	
Conceptualize	Study Concept	Create	Plan	Creating data	Plan and Design	Data acquisition, processing and quality assurance	• Data Management Plan consulting • Planning for storage/backup, metadata creation, archiving, preservation, and access (incorporating best practices/applying standards)
Create or Receive	Data Collection		Collect		Collect and Capture		• Data documentation/metadata support • Local/institutional storage and backup • Support for sharing of data within research team and among collaborators • Data analysis support • Data visualization support
	Data Processing	Describe	Describe	Processing Data		Data description and representation	
Appraise and Select	Data Archiving	Store	Assure	Preserving data	Manage and Preserve	Repository services/preservation	• Policy/legal guidance for data retention • Value assessment (uniqueness, reproducibility, cost to collect/recreate) • Additional metadata creation/transformation support • Assistance locating appropriate repository/archive • Archiving and preservation system(s) • Quality control • Preservation metadata support
Ingest							
Preservation Action							
Store			Preserve				
Access, Use, and Reuse	Data Distribution	Register, Identify	Discover	Giving access to data	Release and Publish	Data dissemination	• Discovery metadata support • Services for exposing metadata • User interface(s) for access and discovery • Assignment of persistent identifier(s) • Support for data reuse/reanalysis • Support for data integration across collections • Advanced querying capabilities (geospatial, temporal, etc.)
	Data Discovery	Discover			Discover and Reuse		
	Data Analysis	Access	Analyze	Analysing data			
			Integrate	Re-using data			
Transform	Repurposing	Exploit					• Data migration • Subsetting capabilities

Appendix D. Data Management Task Force Survey Report

June 2012

The Data Management Task Force (DMTF) members decided to gather broad information through an online campus-wide survey about researcher data management. The short survey was selected in part to complement the Libraries' strategic initiative in support of E-Research, which used targeted interviews to gather similar information about data management at CU Boulder.

The survey was sent to the campus community in January 2012 via the Faculty and Research e-Memo list consisting of 4,411 names, the Buff Bulletin and targeted e-mails to individuals. There were 148 complete responses. While this is a low response rate, it could nonetheless reflect the relatively small number of people currently on campus who are sufficiently versed in data management practice to be able to respond. The twenty-two-question survey was designed to communicate about the DMTF to campus researchers, raise awareness about important questions related to data management, gather information about the data and data management practices across the institution, and offer an opportunity for respondents to give feedback about services and priorities..

The survey responses help to describe the current state of data management activities across local research disciplines. The various survey questions gave information about:

- Respondents and their research areas
- Types of research data generated
- Current storage amounts and projected growth
- Length of time research data will need to be accessible
- Data and metadata formats for storage
- Maintenance of data or metadata documentation
- Implementation of formal data management plans
- People managing data
- Storage and backup technology being used
- Proportion of data that is sensitive, confidential, or proprietary
- Interest in different types of data management services

The survey was written with broad language to make sure respondents from all domains could reply. Given the diverse nature of respondent's roles, departments and research areas, the survey instrument's responses can be applied statistically to a large majority of the university's population of researchers. All schools and colleges at CU Boulder were represented in the responses. More specifically, the scope of the responses includes 38 departments, 6 institutes, 5 centers and 2 programs representing 337 research areas.

Responses also came from Accounting and Business Support, Athletics, CU Museum, Law Library, Museum of Natural History, Office of Information Technology, University Information Systems and University Libraries.

Based on the sample, the majority of respondents at University of Colorado do not have a clear understanding of how best to manage their data in an easy and effective manner.

The following is a summary of findings from the survey:

1. Across campus, respondents need assistance with data management.

An overwhelming percentage of respondents stated that they not only do not have data management plans and/or metadata but that they would like help with planning activities and assistance creating and maintaining data management plans and metadata. With a representation of nearly every academic discipline across the campus, it seems that there is a large demand for education and assistance with regard to data management.

2. An individual CU Boulder researcher will use many different file types and data types for research data.

No single respondent listed less than four different file and/or data types in their responses to the survey instrument. A total of close to 150 unique file types were listed as relevant and important across all responses and in all cases respondents use at least two different types of data such as documents, spreadsheets, digital media, etc.

3. The total data stored by researchers at CU Boulder varies regardless of research area or role.

There was no statistical probability of seeing higher or lower amounts of data across departments and institutes. In other words, in every department or institute across campus there are people with large amounts of data and people with smaller amounts of data. In any given area of campus people are likely to have more than 1MB and less than 10TB of data.

4. A majority of respondents collected the total of their stored data over a period of three years.

While there were respondents on either the short or long end of the time scale, more than 50% of respondents stated that their data was collected in about 3 years.

5. Many of the researchers at CU Boulder need access to their data for at least five years.

Similar to the question above, respondents stated overwhelmingly that they would need access to their data for 5 years. In some cases researchers needed more or less data access, but on the average 5 years was the amount of time that a researcher would be likely to use the data they presently have stored.

6. Most respondents regardless of research area or department do not maintain metadata or data documentation.

There was no statistical difference between departments in their responses to these questions. Only 36% of respondents stated that they do have metadata and these numbers are consistent for all departments across the campus.

7. Few researchers have a data management plan and this is true across all reported research areas and departments.

There was no statistical difference between departments in their responses to these questions. Only 24% of respondents stated that they do have data management plans and these numbers are consistent for all departments across the campus.

8. The vast majority of respondents at CU Boulder manage their own research data.

An overwhelming number of respondents stated that they themselves are responsible for management of their data. A considerably smaller number of respondents stated that others within their department or others outside their department or the university have a hand in managing their data.

Respondents and Research Areas

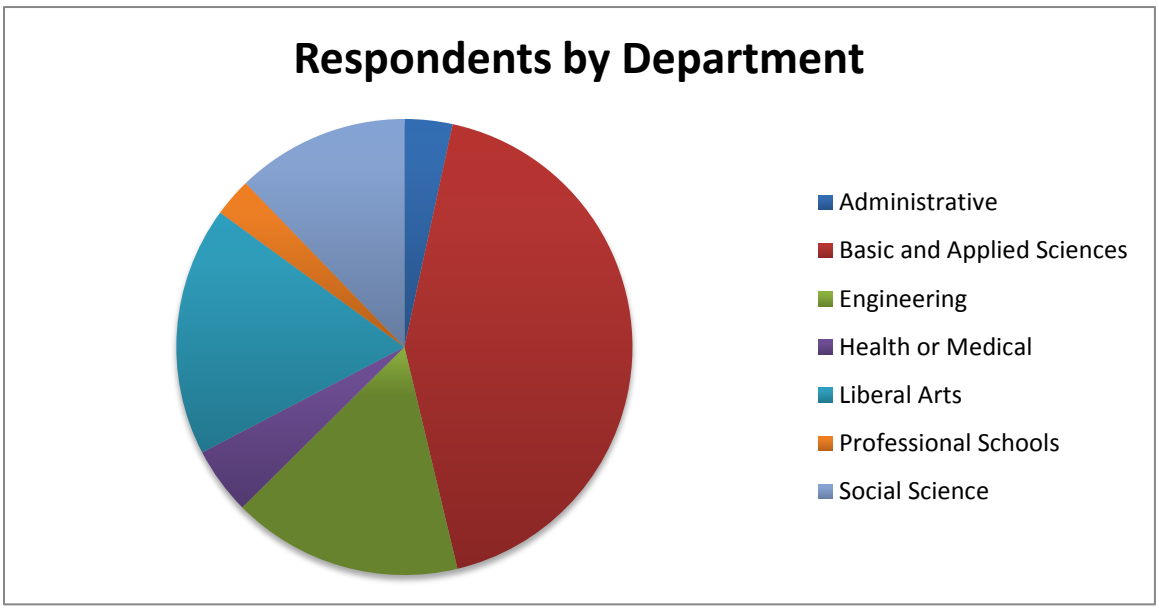
Given the diverse nature of respondent roles, departments and research areas, the responses to the survey instrument can be applied statistically to a large majority of the university's population of researchers.

Table 1: Survey Participation by role and department

Department Category	Admin/Staff	Faculty	Student	Total	Percent of Total
Administrative	2	3	0	5	3.5%
Basic and Applied Sciences	10	43	7	60	42.6%
Engineering and Math	5	17	1	23	16.3%
Health or Medical	2	4	1	7	5.0%
Liberal Arts	1	20	3	24	17.0%
Professional Schools	0	4	0	4	2.8%
Social Science	4	12	2	18	12.8%
Total	**24**	**103**	**14**	**141**	**100.0%**

As seen above in Table 1, faculty constituted the largest category of respondents. Faculty in Table 1, refers to those who self-identified as tenured, tenure track, instructor, lecturer, and researcher.

Chart 1: Respondents by Department



*Applied Sciences: Biology, Chemistry, Physics as well as variations and associated institutes
**Engineering includes Applied Mathematics

Research Data Generated

Each respondent indicated the various types of data generated or used in their research. The data types ranged from simple text data like e-mails and note style documents to complicated simulation or instrument-generated data. Below, in Table 2, the various types of data are labeled along with their frequency.

Table 2: Frequency of Data Types

Type of Data	Frequency of Type of Data	Percent of Total
Computer Programs	101	9.2%
Sensors/Instruments	92	8.4%
Experimental	77	7.0%
Fieldwork	55	5.0%
Lab Notes	53	4.8%
Images	64	5.8%
Web-sites	84	7.7%
Blogs	36	3.3%
E-Mail	91	8.3%
Digital Audio/Video	45	4.1%
Documents	126	11.5%
Spreadsheets	108	9.8%
Databases	93	8.5%
Simulation	72	6.6%
Total	**1097**	

Further statistical analysis was conducted to see if data types cross-tabulated against department/research type. The goal here was to determine if different departments were more likely, statistically, to generate or use certain types of data than others.

The analysis showed that Applied Sciences and Engineering were more likely to report the use of lab specific data types (sensor data, simulation data, experimental data, computer generated data) while professional degrees (Business and Law) and Social

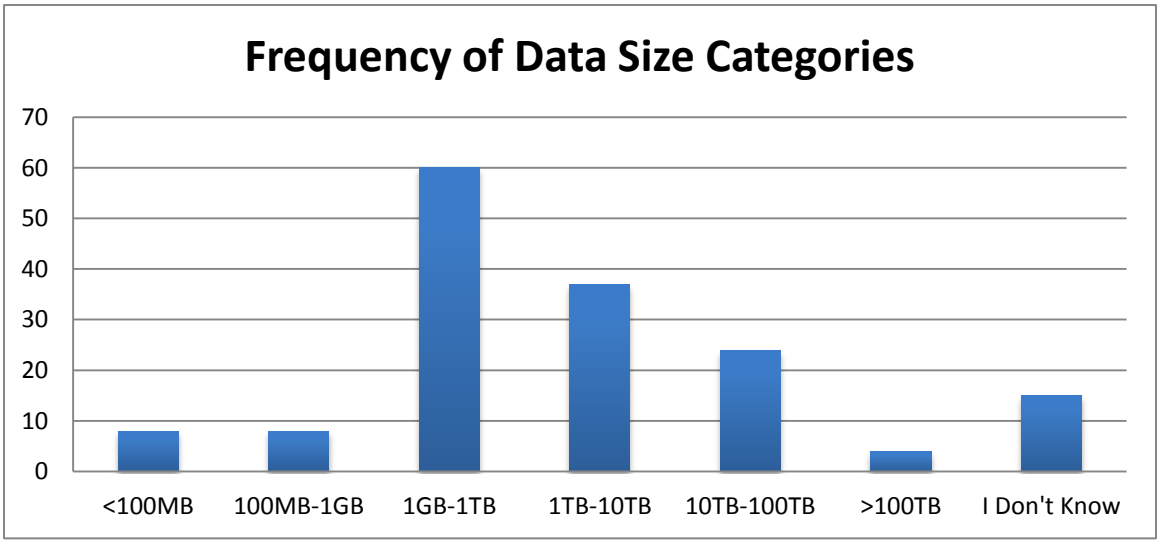
Sciences were more likely to have data related to communication (blogs, websites, and digital media).

Using Chi-Squared statistics related to cross tabulation and Phi and Cramer's V as a statistic related to importance (amount of variability explained as opposed to statistical significance) all data types were analyzed versus department. For examples of the aforementioned statistics and explanations, please see Appendix 1.

Current Storage Amounts and Projected Growth

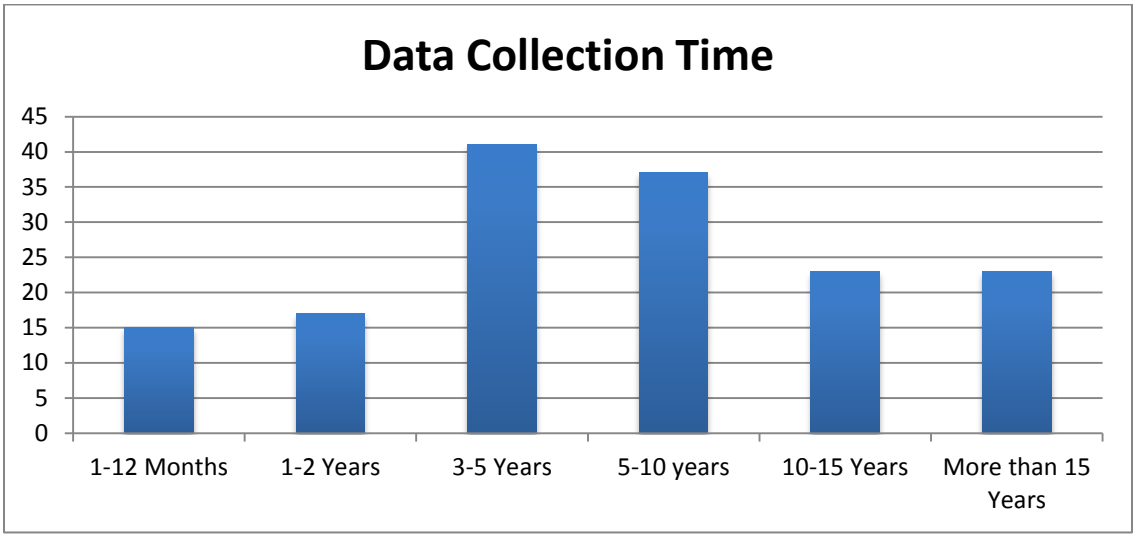
The majority of respondents store between 1 GB and 10TB of research data. Using the cross tabulation technique mentioned above, it was found that Basic and Applied Sciences were most likely to have large data amounts, greater than 1TB.

Chart 2: Frequency of Data Size Categories



Most respondents accumulated the research data they store over 3 years. Using cross tabulation techniques, it was found that there was no significant difference between type of research or department and the amount of time it took to collect the data. This means that across the university, regardless of the researcher's role or research discipline data is collected over varied time periods.

Chart 4: Data Collection Time

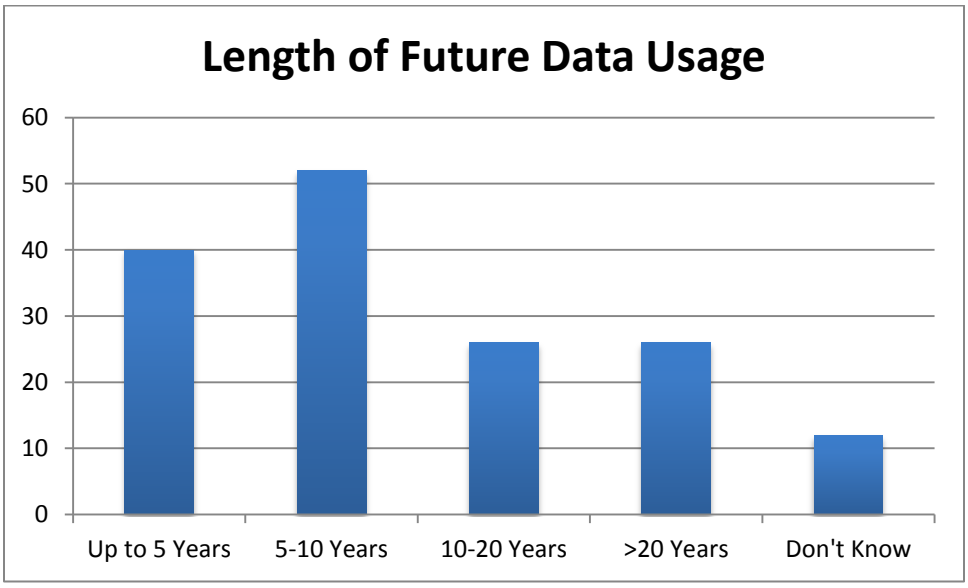


Also, all respondents stated that they expected their data to increase or stay the same in size through the foreseeable future.

Length of Future Data Accessibility

Only slight differences were found when statistically comparing length of future data usage to department or research type of respondent. In most cases professional degree faculty was less likely to need data long term while other academic fields were more likely to need their data for longer periods of time.

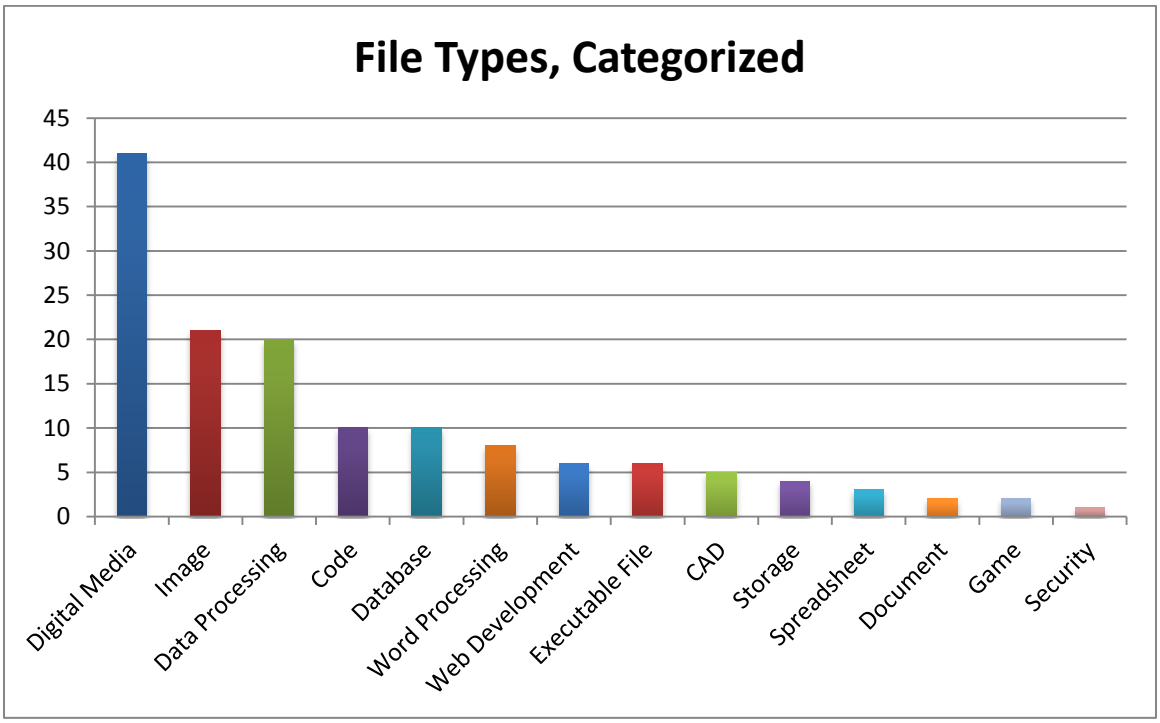
Chart 5: Length of Future Data Usage



Data and Metadata Formats for Storage

Respondents indicated 139 different data and metadata formats used for storing data. This was associated with the different types of data and metadata used, see Chart 6 below.

Chart 6: File Types, Categorized

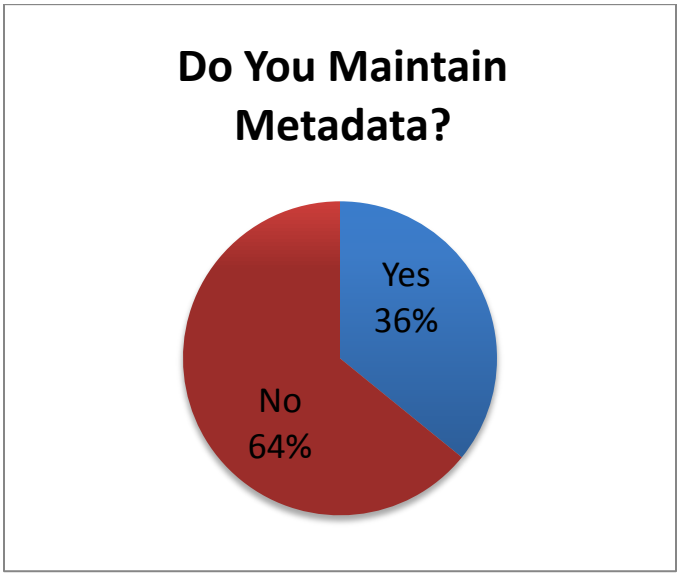


(Category Definitions: *Digital Media*: All files associates with video or audio stored digitally; *Image*: All files that store photographs or images; *Data processing*: All file types used for large data or mathematical modeling such as MATLAB; *Code*: File that contain only code used in other programs; *Database*: All database files such as Access or Oracle; *Word Processing*: Document files such as WORD documents; *Web Development*: HTML, XML or similar files used on the web; *Executable Files*: File that, when clicked on, execute a process; *CAD*: Computer Aided Design files; *Storage*: Flat files that store data such a compressed ZIP files; *Spreadsheet*: Excel or similar files; *Document*: PDF or similar files; *Game*: File associated with digital games; *Security*: File associated with security r virus screening software.)

Maintenance of Metadata or Data Documentation

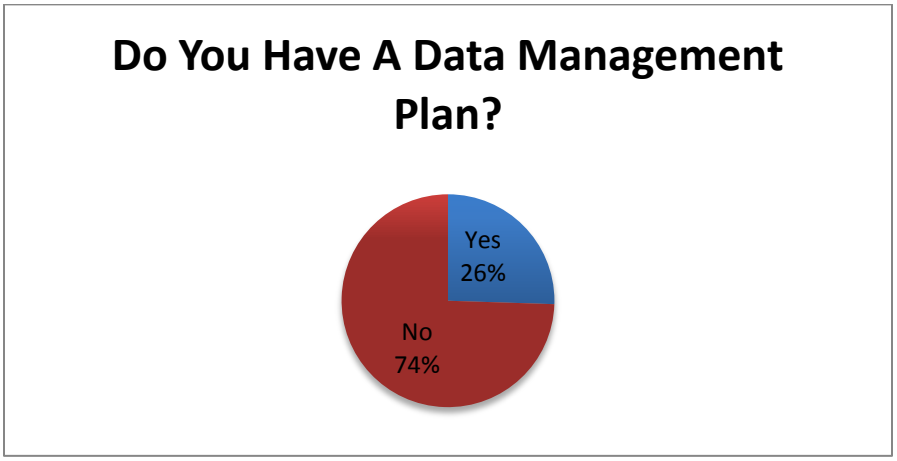
Across the University there was no significant difference in percentage of people maintaining metadata. Regardless of respondent's department or research area, they were more likely to not maintain metadata (64%).

Chart 7: Do You Maintain Metadata?



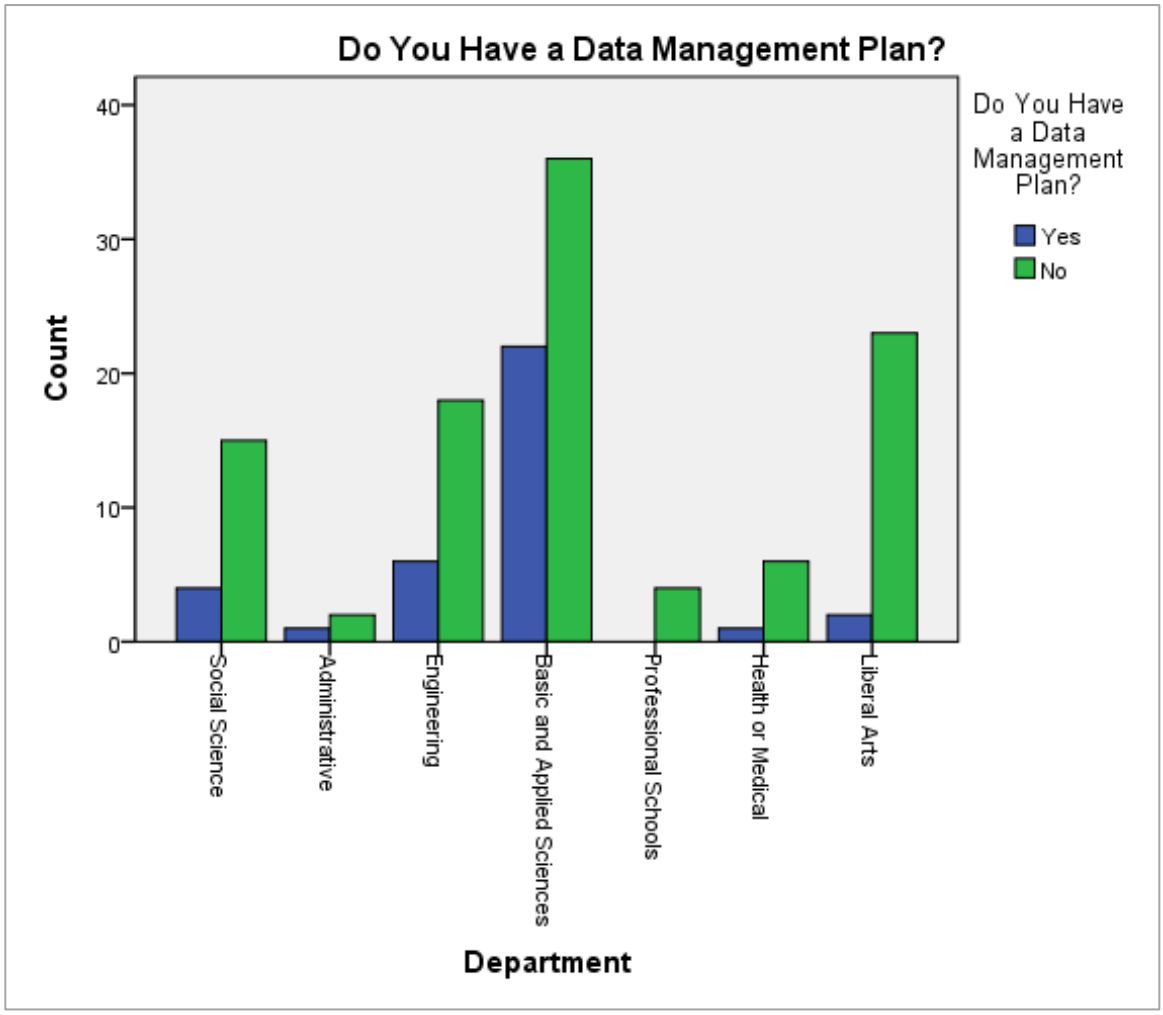
Furthermore, respondents, regardless of their department or research field, responded, approximately 75% of the time, that they did not have a data management plan. This is a very significant result because it proves that all departments need education on both the necessity of data management planning and how to create effective data management plans.

Chart 8: Do You Have A Data Management Plan?



Implementation of Formal Data Management Plan

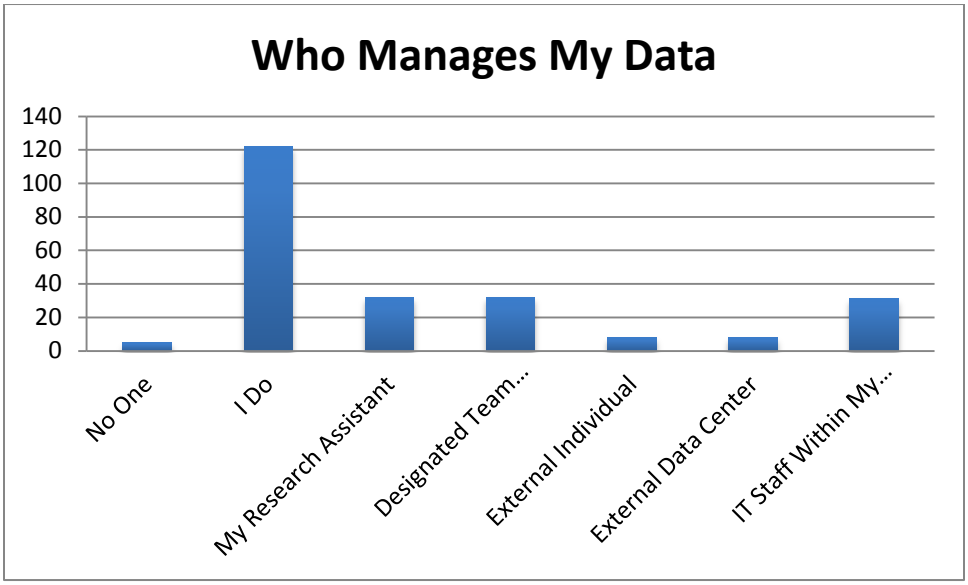
Chart 9: Data Management Plans by Department



The differences shown above were found to be statistically insignificant which means that differences seen are purely due to sampling error. On the average, respondents are significantly less likely to have a data management plan.

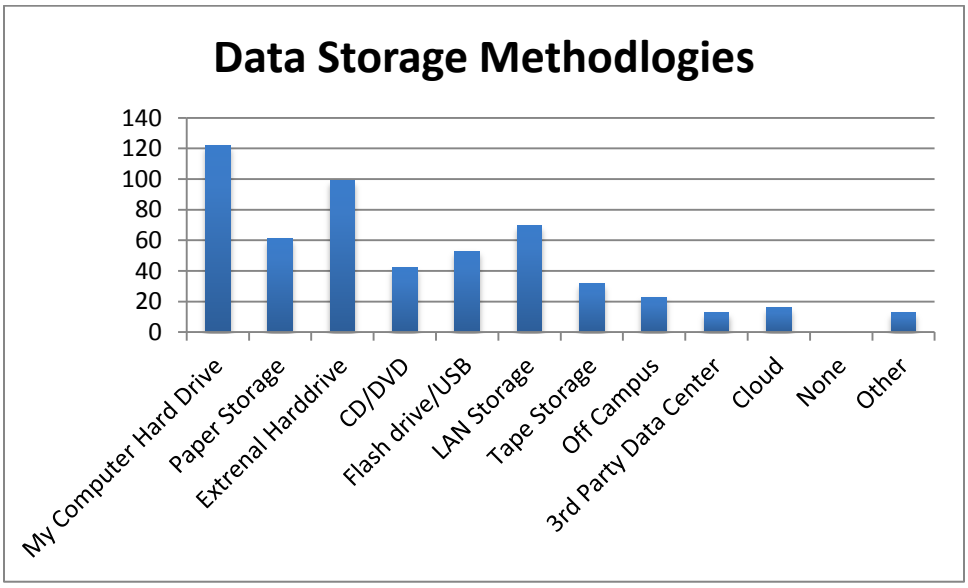
People Managing Data

Across the University the vast majority of respondents stated that they, themselves, manage the data that they store. This statistic allows us to focus our initial education on individuals who are responsible for managing research projects and data.

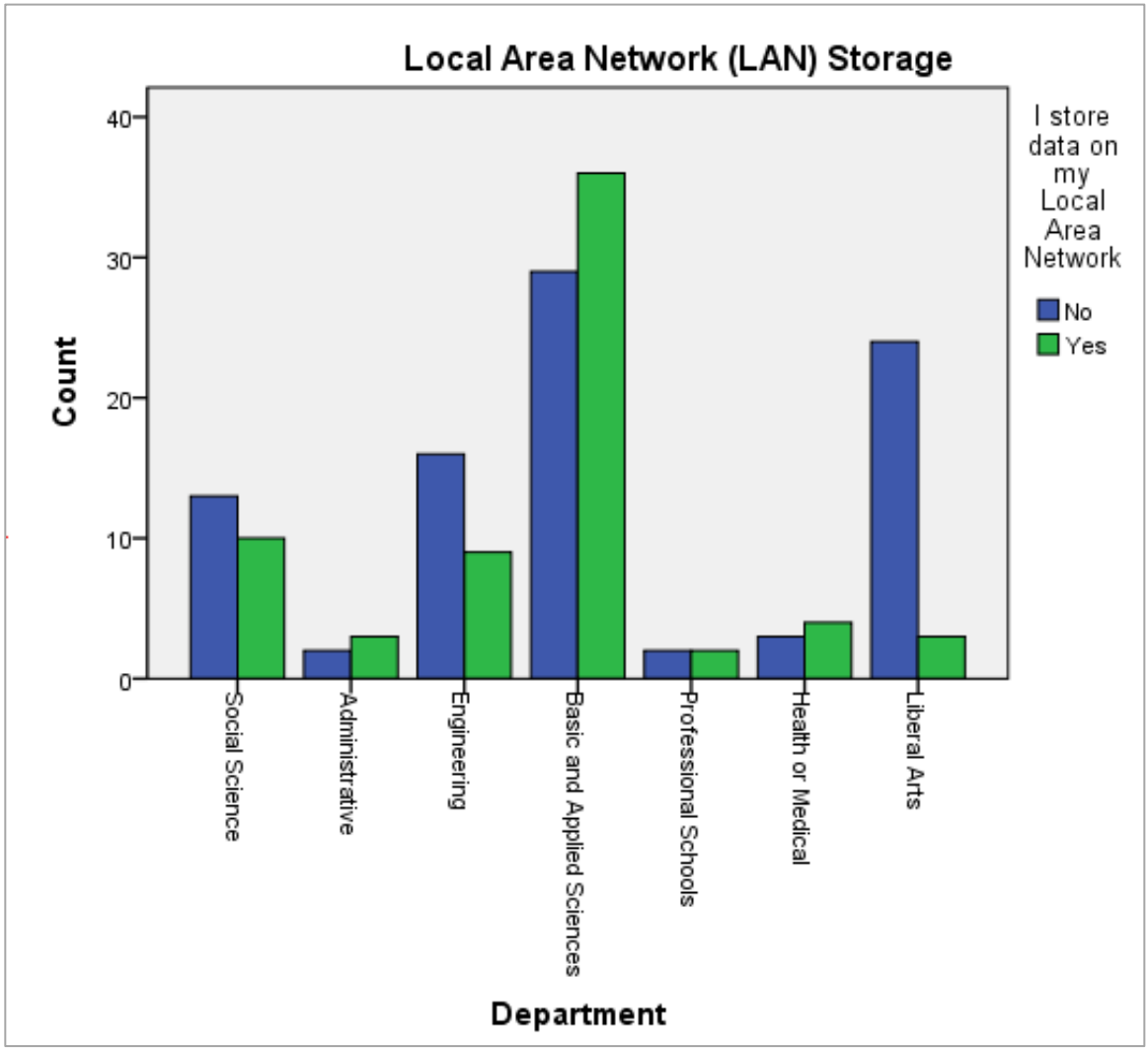


Storage And Backup Technology

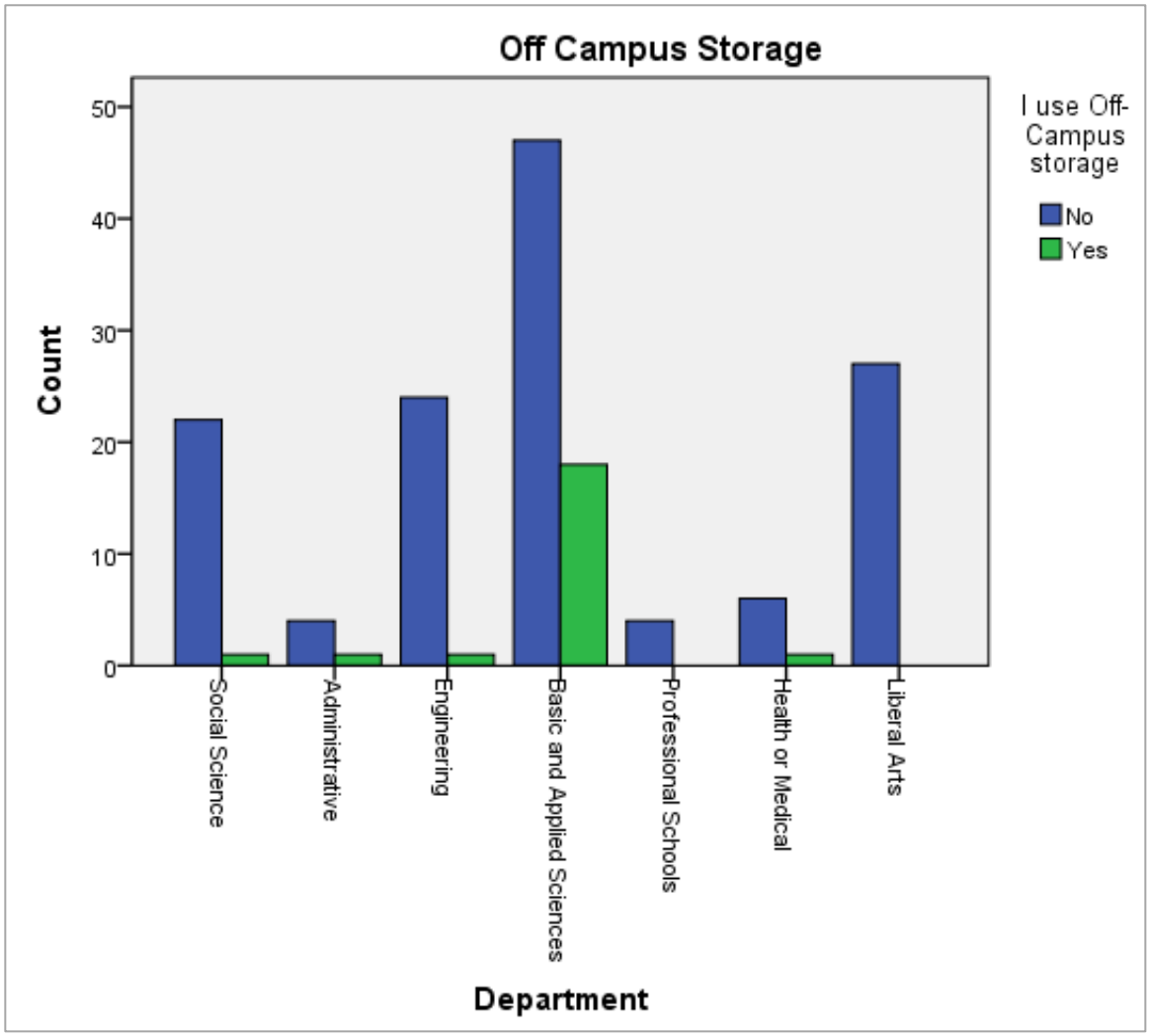
A wide variety of different techniques for storing data are used across the University. For the most part these techniques are reported at the same levels for all departments:



The only interesting areas that differ from the numbers seen above are LAN storage where hard sciences (Physics, Biology, Chemistry, etc.) were more likely to use LAN storage and Liberal Arts were less likely:

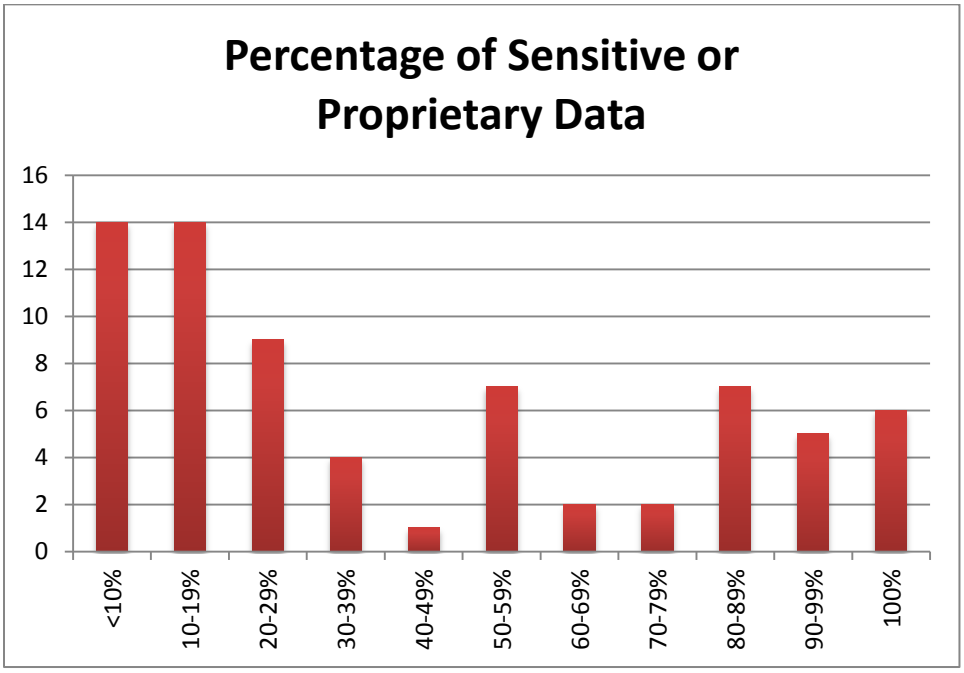


Also of note, off-Campus storage is only used in significant numbers within hard sciences while all other departments had little to no reported usage:



Proportion of Sensitive, Confidential or Proprietary Data

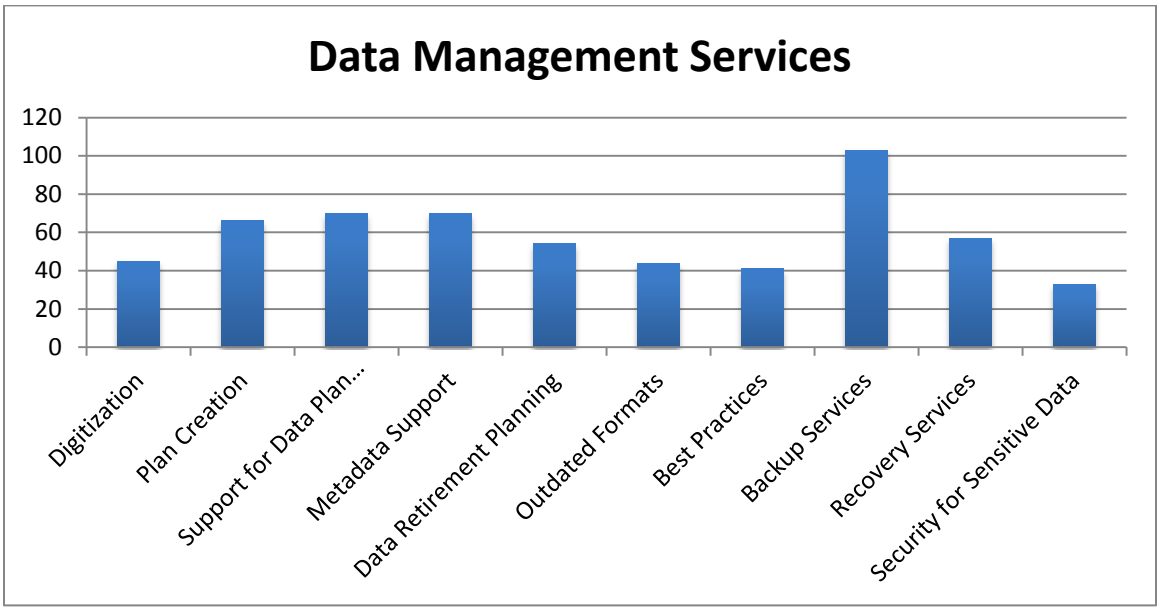
A wide variety of answers were reported with regard to percentage of total data that could be considered sensitive, confidential or proprietary. There was no pattern, statistically speaking, to link amount of proprietary data with department or area of research.



The difficulty with such a wide spread of results is that it means that using inferential statistics to target which areas are more likely to have proprietary information is impossible. For education purposes it should be assumed that any department across the University is equally likely to have researchers with a high quantity of sensitive data.

Interest in Data Management Services

Respondents were given the opportunity to express interest in possible, future data management services.



Using ANOVA to compare response rates, we were able to find 6 distinct groups that services fell into six distinct groupings the analysis results are below:

ANOVA table:

Tests of Between-Subjects Effects

Dependent Variable:Interest in Service

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
ServiceType	24.645	9	2.738	12.272	.000
Error	330.242	1480	.223		
Corrected Total	354.887	1489			

Results of Post Hoc test:

	Subsets				
Type of Data Management Service	1	2	3	4	Point Estimate
Security for Sensitive Data	0.22				**22.0%**
Best Practices	0.28	0.28			
Outdated File Types	0.30	0.30			**29.0%**
Digitization	0.30	0.30	0.30		
Data Exit Planning	0.36	0.36	0.36		
Recovery	0.38	0.38	0.38		**34.7%**
Creation Support		0.44	0.44		**44.0%**
Managing Support			0.47		
Metadata Creation			0.47		**47.0%**
Storage and Backup				0.69	**69.0%**
Sig	0.10	0.07	0.07	1.00	

In the table above, the far right column signifies the percentage of respondents who responded that they were interested in the service in question. Backup and Storage services had, by far, the largest response rate with 69% of respondents saying they would be interested in further help if the service were offered in the future.

At the other end of the spectrum Security Services for sensitive data is the service that respondents were least interested in with a response rate of 22%

In Conclusion

The variability in data and file types across research discipline along with the differences in standards, workflows, instruments and software makes research data management complex. To more completely understand the needs, targeted information gathering will need to be done as resources become available.

Survey Appendix 1: Cross-tabulation Examples and Explanations

Using the Chi-Squared statistic all data in the survey was analyzed using the technique of cross tabulation. This statistic will tell the user whether or not there were differences in how respondents answered a question depending on another, second, question response.

For instance, if we had the question, “Please enter your school, department, or researcher institute:” and wished to know if there were any differences in how respondents answered the question, “Please specify all types of data and metadata generated or used in your research. Select all the following that apply...-Data automatically generated by computer programs,” based on the department they entered, we would use the Chi-Squared Cross-Tabulation technique.

Using the already mention example, here is an example of the statistical output:

Department * Computer Program Automatically Crosstabulation

			Computer Program Automatically		
			0	1	Total
Department	Social Science	Count	9	14	23
		Expected Count	8.6	14.4	23.0
		% within Department	39.1%	60.9%	100.0%
		% within Computer Program Automatically	15.5%	14.3%	14.7%
		% of Total	5.8%	9.0%	14.7%
	NonAcademic	Count	1	4	5
		Expected Count	1.9	3.1	5.0
		% within Department	20.0%	80.0%	100.0%
		% within Computer Program Automatically	1.7%	4.1%	3.2%
		% of Total	.6%	2.6%	3.2%
	Engineering and Math	Count	8	17	25
		Expected Count	9.4	15.6	25.0
		% within Department	32.0%	68.0%	100.0%
		% within Computer Program Automatically	13.1%	16.8%	15.4%
		% of Total	4.9%	10.5%	15.4%
	Hard Sciences	Count	15	50	65
		Expected Count	24.5	40.5	65.0
		% within Department	23.1%	76.9%	100.0%
		% within Computer Program Automatically	24.6%	49.5%	40.1%
		% of Total	9.3%	30.9%	40.1%
	Business and Law	Count	2	2	4
		Expected Count	1.5	2.5	4.0
		% within Department	50.0%	50.0%	100.0%
		% within Computer Program Automatically	3.3%	2.0%	2.5%
		% of Total	1.2%	1.2%	2.5%
	Medical or Treatment related	Count	3	4	7
		Expected Count	2.6	4.4	7.0
		% within Department	42.9%	57.1%	100.0%
		% within Computer Program Automatically	4.9%	4.0%	4.3%
		% of Total	1.9%	2.5%	4.3%
	Liberal Arts	Count	20	7	27
		Expected Count	10.2	16.8	27.0
		% within Department	74.1%	25.9%	100.0%
		% within Computer Program Automatically	32.8%	6.9%	16.7%
		% of Total	12.3%	4.3%	16.7%
Total		Count	61	101	162
		Expected Count	61.0	101.0	162.0
		% within Department	37.7%	62.3%	100.0%
		% within Computer Program Automatically	100.0%	100.0%	100.0%
		% of Total	37.7%	62.3%	100.0%

The above table represents a calculation of each department's response to the question Yes, or No (1 = yes, 2 = no) do you have automatically generated computer data. The expected count field is very important for calculating whether a department is statistically different from the rest of the departments for the question at hand. The expected count is a large part of the Chi-Squared calculation that can be seen immediately below:

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	22.894[a]	7	.002
Likelihood Ratio	22.910	7	.002
Linear-by-Linear Association	6.134	1	.013
N of Valid Cases	162		

a. 8 cells (50.0%) have expected count less than 5. The minimum expected count is 1.51.

In this case, Chi-Squared has a p-vale of 0.002 which, when compared to our cutoff of 0.05, is statistically significant. This tells us that we have statistically sufficient evidence to infer that there is a difference department to department in how respondents answered the question related to automatically generated computer programs.

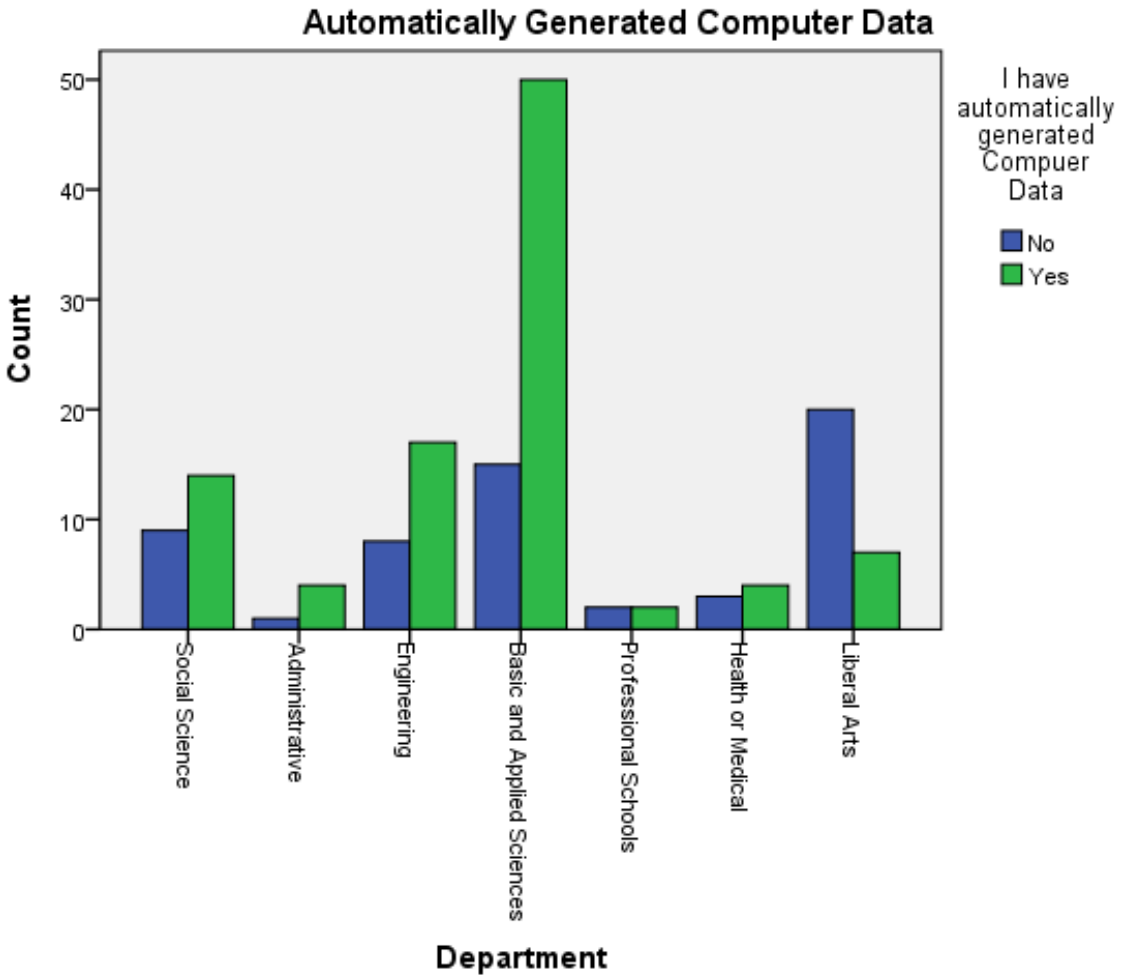
Finally we use Phi and/or Cramer's V to tell us how powerful, e.g. important, the relationship between department and response to automatically generated computer program data is.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.381	.001
	Cramer's V	.381	.001
N of Valid Cases		156	

With a Phi or Cramer's V of 0.381, we can state that there is a moderately strong relationship between the two responses.

The easiest way to interpret this statistic is to look at a bar chart of the information we are analyzing:



The bar chart above shows us that most departments have a 60% yes response to the automatically generated computer data question. But, the reason we have significance with Chi Squared lies in Liberal Arts and Hard Sciences. Hard Sciences relies, more than most departments, on automatically generated computer data while Liberal Arts has a reverse response and relies even less than most departments.

In the case of this particular survey, this sort of analysis is very helpful because it allows us to focus our resources on those departments that most use a certain type of data. The analysis also tells us, when there is no significance, that across the university, all departments have the same likelihood of using a type of data.

Appendix E. *The Strategic Initiative for Research Data Support and Services at the University of Colorado Boulder Libraries*

(Report from CU-Libraries' Participation in the 2011-2012 Association of Research Libraries E-Science Institute)

Executive Summary

Universities across the country are grappling with how to increase the value of research data. At the same time research libraries nation-wide are undergoing many changes to address digital and data curation needs. The complexities regarding data include topics such as changing federal mandates, university policies, intellectual property laws, use of new and disruptive technologies, organizational restructuring, and more. In order to address new mandates, university leadership must build support by communicating to campus constituents, taking advantage of cross-functional expertise, clarifying expectations, and developing tools to support data sharing and preservation. Considering the broader impacts of managing and sharing the data for the good of the institution is an essential part of the "data value cycle". To address this issue, Research Data Services at CU was conceived to represent the cross-functional collaboration necessary to solve these complex data management, sharing, and preservation needs. Specifically, Research Data Services is a partnership between the Office of Information Technology (OIT), the Office of the Vice Chancellor for Research, Research Computing (RC), and the University Libraries. Research Data Services will provide a venue for all of the partners involved to work together, each with their own strength, to support researchers in the earlier parts of the data life cycle and the campus in the latter. This document focuses on the University Libraries' role in this partnership.

Background

In order to explore the complexities surrounding data management issues and to identify the role for the Libraries at CU Boulder, the University Libraries participated in the Association of Research Libraries (ARL) E-Science Institute. Team members representing CUB were Suzanne Larsen, Interim Associate Dean of Libraries; Barbara Losoff, Science Librarian; and Kimberly Stacey, Research Data Manager with Research Computing. The Institute was a six-month process that involved information gathering through questionnaires, readings, interviews, and webinars. The culmination of the Institute was a capstone event in January 2012. E-Science teams from participating universities met face-to-face to share information. The outcome is this strategic initiative which addresses the specific environment at the University Libraries at CU Boulder.

Much of the work at the Institute capstone was done through collaboration with other teams. Our partners at various times were: University of Oregon, Cornell, Rice, and UCLA. We also had conversations with attendees from: UCSD, University of North Carolina - Chapel Hill, University of Chicago, University of Illinois, Virginia Tech, and Purdue. These interactions, the planned exercises at the capstone, our self-assessment, questionnaires, interviews, and SWOT analyses form the basis for this strategic initiative. In addition, these efforts complement the data gathering efforts by the Data Management Task Force (see charge in Appendix 3) in their

campus-wide survey and ongoing work.

I. **Summarized Self-assessment (see Appendix 1)**

Overview of campus research and technology organization and support

CU Boulder is a Tier 1 research university receiving over $359 million in new sponsored research projects in fiscal year 2010-2011. These awards, tracked by the Office of Contracts and Grants (OCG), include grants in the natural and physical sciences, social sciences, arts, humanities, space sciences, and engineering. The Institutes, in particular CIRES ($61 Million), LASP ($55 million), JILA ($22 million), and IBG ($12 million), had the highest awards. The highest departmental awards were Chemistry and Biochemistry, Physics, MCDB, and Psychology and Neuroscience. Research at CU is highly collaborative which is due in part to the number of national labs located nearby. For example, CIRES has strong ties to NOAA and NCAR, the Physics Department is closely connected to NIST, and several departments and institutes are aligned with NREL.

Research support at CU Boulder is directed by the Office of the Vice Chancellor for Research. The Office of Contracts and Grants (OCG), which reports to the VC for Research, is responsible for the administration of sponsored research agreements and funding. OCG assists faculty, staff, and students in obtaining and managing external support for their sponsored projects, while ensuring compliance with sponsor policies and procedures and protecting the interests of the University and the State of Colorado. The Director of Research Computing (RC) reports to both the VC for Research and the Associate Vice Chancellor for Information Technology/CIO. Research Computing is tasked with developing cyberinfrastructure and support for campus researchers. Research Computing offers support with writing data management plans and provides access to high performance computing resources, including parallel processing and a data intensive network. The Office of Information Technology provides central services and resources for enterprise level and administrative support. In addition, OIT has a group that supports academic and educational goals for faculty and students.

Traditionally, the CU Libraries has supported campus research through the provision of scholarly journals, monographs, and documents, as well as services such as interlibrary loan. Librarians consult with faculty regarding resources to support their research, locate hard-to-find citations, and assist with literature reviews. However, the role of libraries in support of the university research mission is changing nationwide. By leveraging faculty and staff members' collective experience of working closely with researchers and providing long-term access to digital materials, the University Libraries can play a significant role in the data management process. In anticipation of that role, individuals with expertise in metadata and other emerging areas of librarianship directly related to research data management are now members of the Libraries

faculty. This also includes subject specialists with knowledge of a wide range of research disciplines. In addition, CUB librarians are meeting and communicate regularly about research data management issues. Several Libraries faculty members consider this to be a primary research interest.

II. Summary of SWOT (see Strengths, Weaknesses, Opportunities, and Threats, Appendix 2)

- The Vice Chancellor for Research supports the establishment of Research Data Services and sees a role for Libraries in describing data to make it discoverable (S1). The Dean of the Libraries strongly supports the Libraries participation in data management services as part of a campus-wide partnership that includes RC and OIT (S2). He is a member of the Boulder Campus Cyberinfrastructure Board (BCCB). BCCB is the Board that sets policy for research computing on the campus.
- Dean Williams envisions the Libraries becoming a hub in this partnership. Because data is an institutional asset, how it is managed and shared is a benefit to both the University and the community (O1). Formalizing and funding these partnerships will aid researchers by providing structured data management services (T3) which will offset the need to use data storage outside a managed, shared environment (T2).
- The Libraries capacity for skill development in data services through campus partnerships (S6), the hiring of a Metadata Librarian (S3), and the acquisition of DigiTool for creating an institutional repository (O3), suggests that the Libraries are ready to respond to the Campus' growing need (O6) for data support and services.
- The Libraries have an opportunity to provide assistance to individuals in departments, Institutes and Centers by developing tools and services for data management in cooperation with Research Data Services (O5). Librarians with subject expertise can assist faculty and graduate students with data management literacy and support. Additionally, the OCG, which has no mandate or funding to support PIs with their data management plans (W1), supports the plan to connect PIs with Research Data Services (O7).
- Data management at CUB will only be successful with the support from Campus administration. Currently, there is only one fully-funded position dedicated to data management services (W7). The University needs to evaluate and fund the appropriate level of support for data management services. In supporting Research Data Services, the University has the opportunity to demonstrate to major funding agencies that CUB is making institutional cultural shifts in order to comply with data management mandates (O11).

III. Potential roles and services for University Libraries

A. Data management / curation

- Identify best practices for data collection, sharing, and preservation.
- Create collection development policies for data with criteria to determine what should be saved and for how long. These policies should address data reuse.
- Provide one-on-one consulting for creating Data Management Plans (DMP), helping researchers reduce the time spent on this activity.

B. Data Analysis

- Provide support and purchase software for statistical analysis and interpretation of data.
- Help researchers identify and locate quality sources of data from the Libraries' collections and online data archives/repositories.

C. Metadata / ontologies

- Consult with researchers to describe data and create workflows for metadata provision
- Use registries and other resources to identify schemas and ontologies for appropriate disciplines or data types

D. Outreach / support

- Act as a liaison, when necessary, by directing researchers to specialized resources (e.g., disciplinary data archives) or other Research Data Services partners.
- Brand research data so it can be identified with CUB in order to demonstrate value to the University and community.
- Help with data governance through guidance during system-wide policy development and implementation.
- Actively participate in helping OCG to become an integrated hub for referrals between Libraries, Research Data Services, and researchers.
- Provide outreach and education about the value of good data management practices (e.g., using standards/ontologies) to PIs and graduate students involved in data collection.

IV. Assessment of opportunities

Currently, data management services are supported through individual planning and local efforts between the Libraries and RC. This results in a disjointed provision of services, thereby contributing to silos of information with no potential for scalability. If the campus does not move forward in planning and initiating centralized Research Data Services, the workload will become unmanageable and there is a real risk of losing the trust and support of researchers and potentially valuable data.

The Libraries needs to implement solutions soon, or run the risk of being marginalized and under-utilized by the research community. In addition, if the Libraries are not seen as a valuable partner, there is a risk that data services will be monopolized by entities that see data management exclusively as an issue of storage capacity and do not recognize the potential value lost by neglecting curation and preservation activities. The Libraries has the opportunity to work with OCG as a conduit to establish contact with researchers during the initial proposal submission process to provide data management services. If the Libraries do not demonstrate their value as a partner, in cooperatively developing data management services, that connection to OCG will be lost. The Libraries must take a leadership role system-wide, collaborating with the sister institutions ensuring that services are not redundant or overlapping.

Risks of Failure

One risk in pursuing data services and initiatives is that the Libraries will be unable to provide adequate support for data management activities. In order to successfully support these activities, the Libraries will need to provide training for current staff and hire new personnel with appropriate skill sets. The Libraries will need to reallocate positions, which will impact other library services. One risk is that these changes may not be accepted by the library personnel, while another risk is that changes to the service model may not be accepted by the Campus at large. In addition, there is also the risk of being too successful and not having the resources to make the services scalable. Either way, the Libraries, through careful planning, is committed to supporting data management services.

V. Recommendations

- STAFFING
  The Libraries need to realign their mission to support data management services. Building on the internal reorganization, the Libraries can begin to assign percentages of time or specific people for data management activities. New positions will need to be created, and hiring for those positions will need to be supported by the Campus at large. The Libraries should explore options such as temporary hires with the needed skills sets in order to move forward in a timely manner. The Campus at large also will need to assign personnel resources to Research Data Services.

- TOOLS
  The Libraries recently acquired DigiTool, a management system and repository for digital objects. Currently, the storage dedicated to DigiTool is minimal and is not a viable option for housing large datasets. DigiTool could be used to archive final, completed, smaller datasets. However, without versioning capabilities and other advanced data management features, DigiTool would not be an appropriate solution for complex and/or dynamic data objects. The Libraries needs to clearly define and communicate to the

campus what types of data are suitable for archiving in DigiTool. When possible, the Libraries should ensure that DigiTool provides metadata and links to data housed elsewhere in order to provide a single catalog of data produced by campus researchers.

In addition, the Libraries should work with other institutional partners to develop a Research Data Services website in order to serve as a single point of entry for researchers seeking resources for data management.

- INSTITUTIONAL PARTNERS
  The Libraries needs to continue to work with RC to create a single unit that supports data management, Research Data Services. The Libraries should continue to explore partnering with OCG to provide a 'pipeline' to researchers and become involved at the beginning of the research cycle. Within the University System, the Libraries should develop a system-wide initiative with the sister institutions for data services.

- EXTERNAL PARTNERS
  The Libraries have long-standing partnerships with NREL, NCAR, NOAA, and NIST, all of which are developing data services. The Libraries, mirroring the researchers, must seek out the expertise within the national lab libraries and develop joint data services. The Libraries should maintain connections with cohorts from the ARL e-Science Initiative so as not to reinvent the wheel.

In conclusion, it is our hope that the information in this report will generate the necessary support to move forward with these recommendations. Action items, detailed recommendations, and budget requests can be developed in collaboration with the Data Management Task Force discoveries when additional resources are available.

Libraries Report Appendix I: Self Assessment

A. Research Support Structure

The Vice Chancellor for Research, Stein Sture, oversees research support at the University of Colorado Boulder. The Office of Contacts and Grants reports to the VC for Research and oversees the submission and compliance for research grants, Randy Draper is the Director of OCG. The Director of Research Computing is Thomas Hauser who reports to both the Vice Chancellor for Research and the Associate Vice Chancellor for Information Technology and CIO, Larry Levine. Research Computing is tasked with providing research services, enhanced data intense network, storage, and high performance computing resources for researchers.

B. Technology Support Structure

Technology support is tiered at CU Boulder. There are central services and resources provided through the Office of Information Technology (OIT) for enterprise level and administrative support. The OIT also has a group supporting academic and educational goals for faculty and students. These groups all report to the CIO. The Research Computing group reports to the same CIO but is a separate entity with dotted lines to groups in OIT when collaboration is the best solution for campus. There is more overlap in terms of support for the network and supercomputer facilities than is obvious when viewing an organizational chart. Support for technology is also very distributed in the Institutes and Departments. Many departments have their own IT group for support.

C. Research Landscape

The Office of Contracts and Grants (OCG) is responsible for the administration of sponsored research agreements and funding. OCG assists faculty, staff, and students with obtaining and managing external support for their sponsored projects while ensuring compliance with sponsor policies and procedures and protecting the interests of the University and State of Colorado. Additionally, OCG interprets and, if necessary, enforces campus, University, and sponsor policies and procedures; provides training to faculty and staff; and serves as the University's liaison with sponsors and regulatory agencies. OCG works closely with Sponsored Projects Accounting to ensure that all sponsored projects are financially compliant and fiscally sound.

There is no specific e-science support initiative for the Campus. The Libraries has faculty who attend the Vice Chancellor for Research's Research Council and Data Management Task Force. This will position the Libraries to be interactive and reactive to proposals for e-science as they develop. The Libraries and Research Computing have also made a connection with the Office of Contracts and Grants to provide information regarding data management plans and computing support.

D. Level of Research Funding

CU-Boulder is a Tier 1 research university with over $359,million in new sponsored research projects in fiscal year 2010-2011 across the natural and physical sciences, arts, humanities, social sciences, space sciences, and engineering. This is tracked by the OCG. The Institutes, particularly CIRES ($61 million) LASP ($ 55 million), JILA ($22 million) and IBG ($12 million) had the highest awards. The highest individual departmental awards were to Chemistry and Biochemistry, Physics, MCDB, and Psychology and Neuroscience.

Funding Agency	# Awards		Dollars
DEPARTMENT OF AGRICULTURE	5		196,213
DEPARTMENT OF COMMERCE	51		47,976,041
DEPARTMENT OF DEFENSE	83		16,495,806
DEPARTMENT OF EDUCATION	19		4,641,668
DEPARTMENT OF ENERGY	60		19,553,922
DEPARTMENT OF ENERGY LABS (except NREL)	29		1,773,519
DEPARTMENT OF HEALTH AND HUMAN SERVICES	226		49,870,128
DEPARTMENT OF THE INTERIOR	51		3,014,004
NATIONAL AERONAUTICS AND SPACE ADMINISTRATION	260		61,123,457
NATIONAL SCIENCE FOUNDATION	317		65,518,948
SMITHSONIAN INSTITUTE	12		489,355
OTHER FEDERAL AGENCIES	30		3,228,885
Non-Federal agencies:			
JET PROPULSION LABORATORY	30		3,548,976
NATIONAL RENEWABLE ENERGY LABORATORIES	51		2,860,018
OTHER UNIVERSITIES	201		30,221,412
SPACE TELESCOPE SCIENCE INSTITUTE	18		693,145
STATE OF COLORADO (includes state funded universities)	63		7,661,756
INDUSTRY	222		18,519,093
FOUNDATIONS	71		8,221,221
FOREIGN UNIVERSITIES AND FOREIGN INDUSTRY	22		1,355,179
ORGANIZATIONS AND ASSOCIATIONS	124		11,338,842
OTHER	9		827,489

TOTAL AWARDS RECEIVED IN FISCAL YEAR 2011	**1,954**		**359,129,077**

E. Collaboration

Research at CU Boulder is collaborative in nature, primarily because of the number of national labs in Boulder and nearby. CIRES has strong ties to NOAA and NCAR. The research in the Department of Physics and JILA is closely connected to NIST. There are also many ties to NREL.

F. Cyberinfrastructure support

Research Computing offers support with data management plan writing provides access to high performance computing resources including parallel processing and a data intensive network. Additionally, small scale storage is to be expanded very soon to long-term hierarchical storage. Research Computing hosts the Janus supercomputer, a 1368 compute node resource interconnect with QDR Infiniband with approximately 900 TB or high performance storage accessible via a Lustre file system. The management network is 1 Gbps and the cluster connects to the CU data intensive network.

We are member of the Front Range Consortium for Research Computing (FRCRC) which includes NREL, NOAA, NCAR, Mines, CSU, CU Boulder and Wyoming.

Libraries Report Appendix 2: Strengths, Weaknesses, Opportunities and Threats (SWOT) Assessment

A. Strengths

S1. Vice Chancellor for Research supports the establishment of Research Data Services on campus and sees a role for Libraries to describe the data to make it findable.
S2. The Dean of Libraries strongly supports the Libraries participation in data management services for the campus with strong partners in Research Computing and Office Information Technology.
S3. The Libraries have hired a Metadata Librarian.
S4. Research Computing (RC) is laying a foundation for data management in the form of cyberinfrastructure with storage networks and system administration for data management.
S5. Research Data Services is developing tools and services for data management, building on the foundational layer of Research Computing. (we haven't said anything about this before)****
S6. Multiple groups on campus are working in tandem regarding the development of data management policies and services.
S7. The number of funded research grants is exceptionally high and growing at CU Boulder.

B. Weaknesses

W1. The Office of Contracts and Grants (OCG) is under-staffed for the amount of research grant money that is flowing into the University, because of that they are unable to provide additional support for PIs. OCG is not privy to nor do they have time to provide guidance and feedback to the PIs for data management plans.
W2 There is no directive (incentive) for the OCG to implement accountability for PIs in fulfilling the data management plans.
W3. Even if researchers wanted to move grant monies toward research data management support there is no organizational mechanism for transferring grant money toward centralized research data services.
W4. Major funding agencies that are important to CU researchers do not allow line-items for research data support, such as development personal, system administration personal, metadata personal, etc.
W5. For researchers, the push to fund data management plans using their grants, is viewed as subtractive, taking away basic funding for their research.
W6. In general there is an overall lack of funding and dedicated personal for data management and services
W7. The University has only one officially fully-funded person on-campus dedicated to strategizing and building-out Research Data Services. There is a 'dotted-line' to others but it's not official.
W8. Researchers are not informed as to where to go to get help with data needs and

the service offered by RC and the Libraries is only in its infancy.
W9. Technical developers are not available within the campus structure

C. Opportunities:

O1. There is an opportunity for the campus to show value to the community regarding research data created by the University of Colorado through setting-up a mechanism for sharing data.
O2 With the newly created and newly restructured units on campus, the establishment of a new entity for data management will find fewer barriers and have more flexibility.
O3. Research disciplines on this campus are data driven, i.e., they understand the challenges of managing data.
O4. Outreach and education opportunities for data management
O5. Potential for leadership and collaboration within the University of Colorado system.
O6. The Libraries has an opportunity to provide a new service which will only continue to grow and serve the broader campus research community.
O7. OCG is interested in being a bridge between PIs and a resource for data management in order to help PIs.
O8. The acquisition of DigiTool, a digital asset management system that will be used as the platform for an institutional repository in which some types of data (e.g., completed datasets) could be archived.
O9. Opportunity to provide researchers with more time for their research and less time on data management.
O10. Opportunity to work with institutes and laboratories to better understand discipline specific data management practices.
O11. Opportunity to show NSF that CU Boulder is making institutional cultural shifts to comply with their data management mandate.

D. Threats

T1. Weak economy and continued lack of funding for higher education in Colorado
T2 Researchers are using easily attained data storage outside of a sharable managed environment.
T3. Lack of institutional policy from the University mandating researcher to archive their data with the University.
T4. Security concerns

Libraries Report Appendix 3: Original email to potential members of the Data Management Taskforce

Dear _____:

I am pleased to invite you to serve on the Data Management Task Force of the Office of the Vice Chancellor for Research' Research Computing initiative. The following charter or purpose for the task force has been developed:

The task force will work to pull together disparate but critical entities and expertise in the CU-Boulder community and will act as a nexus for leading data management efforts. This task force will make recommendations about the storage and curation of digital data produced in the course of CU-Boulder based research. It will address the roles of individual researchers, departments and institutes, staff, and the university as a whole. A wide array of data will need to be considered (e.g., observational, experimental, clinical, simulation). In addition the task force will evaluate how best to manage

- *Data sets that vary substantially in terms of size*
- *Appropriate security for different data sets*
- *Which data need to be retained and for how long*
- *Governance issues such as data ownership, stewardship, access and sharing*
- *Necessary policies; and complications that might arise through collaborations with other entities*

It will also address storage and maintenance issues in both the short and long term, and potential funding models for each. A standardized cost structure will be helpful to investigators because increasingly, funding agencies require recipients of grants and contracts to provide and implement a data management plan. The task force will provide specific recommendations about how CU-Boulder investigators can respond to NIH and NSF policies, although its mandate is not restricted to particular funding sources, or limited to funded research. Because there are many parts of this discovery process that are not unique or restricted to CU-Boulder, the task force will review policies and practices at other universities as well as the national context in formulating its recommendations for CU-Boulder.

Please let me know via return e-mail if you are able to serve on this task force.

Best regards,

Stein

Stein Sture, Ph.D.

Vice Chancellor for Research
Huber and Helen Croft Endowed Professor
College of Engineering and Applied Science

University of Colorado
Boulder, Colorado 80309-0026
(303) 492-2890 FAX: (303) 492-5777
stein.sture@colorado.edu

Appendix F. Selected Universities with Research Data Management Web sites

California Digital Library (University of California System)
http://www.cdlib.org/services/uc3/datamanagement/index.html

Cornell University
https://confluence.cornell.edu/display/rdmsgweb/Home

Duke University
http://library.duke.edu/data/guides/data-management/index.html

Harvard University
http://isites.harvard.edu/icb/icb.do?keyword=k78759

Johns Hopkins University
http://dmp.data.jhu.edu/

Massachusetts Institute of Technology
http://libraries.mit.edu/guides/subjects/data-management/index.html

Purdue University
https://research.hub.purdue.edu/

Stanford University
https://lib.stanford.edu/data-services

University of California, San Diego
http://libraries.ucsd.edu/services/data-curation/

University of Edinburgh
http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/research-data-mgmt

University of Florida
http://guides.uflib.ufl.edu/datamanagement

University of Michigan
http://www.lib.umich.edu/research-data-management-and-publishing-support

University of Minnesota
https://www.lib.umn.edu/datamanagement

University of North Carolina at Chapel Hill
http://www.lib.unc.edu/datamanagement/

University of Oregon
http://libweb.uoregon.edu/datamanagement/

University of Oxford
http://www.admin.ox.ac.uk/rdm/

University of Virginia
http://www2.lib.virginia.edu/brown/data/

University of Wisconsin-Madison
http://researchdata.wisc.edu/

References

Bowker, G. C. , Baker, K., Millerand, F. & Ribes, D. (2010). Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment. In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International Handbook of Internet Research* (97-117). New York: Springer.

Burton, A. & Treloar, A. (2009). Designing for discovery and re-use: The "ANDS Data Sharing Verbs" approach to service decomposition. *International Journal of Digital Curation, 4*(3), 44–56. doi:10.2218/ijdc.v4i3.124

Crowston, K. & Qin, J. (2011). A Capability Maturity Model for Scientific Data Management: Evidence from the Literature. *Proceedings of the American Society for Information Science and Technology*. *48*. doi:10.1002/meet.2011.14504801036.

Edwards, P. N., Jackson, S. J., Bowker, G. C. & Knobel, C. P. (2007). *Understanding Infrastructure: Dynamics, Tensions, and Design*. Arlington, VA: National Science Foundation. http://hdl.handle.net/2027.42/49353

Hey, T., Tansley, S., and Tolle, K. (Eds.) (2009). *The fourth paradigm: Data-intensive scientific discovery*. Redmond, WA: Microsoft Research. Russell Sage Foundation.

Higgins, S. (2008). The DCC Curation Lifecycle Model. *International Journal of Digital Curation 3*(1), 134–140. doi:10.2218/ijdc.v3i1.48

International Standards Organization (2012). *Space data and information transfer systems: Audit and certification of trustworthy digital repositories*. ISO16363:2012.

Jones, K. (2011). Research360: Managing data across the institutional research lifecycle. Poster presented at the 7th International Digital Curation Conference, Bristol, UK, 5–8 Dec. http://www.dcc.ac.uk/webfm_send/589

Macdonald, S. & Martinez-Uribe, L. (2010). Collaboration to data curation: Harnessing institutional expertise. *New Review of Academic Librarianship*, *16*(sup1), 4-16.

Michener, W.K. & Jones, M.B. (2012). Ecoinformatics: Supporting ecology as a data-intensive science. *Trends in Ecology and Evolution*, *27*(2), 85–92. doi:10.1016/j.tree.2011.11.016

MIT Libraries (2009). What is data? Retrieved from http://libraries.mit.edu/guides/subjects/data-management/what.html

National Research Council (2009). *Ensuring the integrity, accessibility, and stewardship of research data in the digital age*. Washington, DC: The National Academies Press.

National Science Board (2011). Digital research data sharing and management. Retrieved from http://www.nsf.gov/nsb/publications/2011/nsb1124.pdf

National Science Foundation (2011). Grant proposal guide. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpgprint.pdf

Office of Science and Technology Policy, Executive Office of the President of the United States (2012). Fact sheet: Big data across the federal government. Retrieved from http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final.pdf

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS ONE, 2*(3), e308. doi:10.1371/journal.pone.0000308

Schofield, P.N., Bubela, T., Weaver, T., Portilla, L., Brown, S.D., Hancock, J.M., Einhorn, D., Tocchini-Valentini, G., Hrabe de Angelis, M., Rosenthal, N., & CASIMIR Rome Meeting participants (2009). Post-publication sharing of data and tools. *Nature*, 461, 171-173. doi:10.1038/461171a

Star, S. L. & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, *7*(1):111.

Stodden, V. (2009). Enabling reproducible research: Open licensing for scientific innovation. *International Journal of Communications Law and Policy*, (13), 1-25.

Structural Reform Group (2004). DDI Version 3.0 Conceptual Model. Data Documentation Initiative.

UK Data Archive (2012). Research Data Lifecycle. http://www.data-archive.ac.uk/create-manage/life-cycle

University of Colorado at Boulder (n.d.). Flagship 2030: Serving Colorado, engaged in the world. Retrieved from http://www.colorado.edu/flagship2030/downloads/CUFlagship.pdf

Van House, N. A., Bishop, A. P. & Buttenfield, B. P. (2003). Introduction: Digital Libraries as Sociotechnical Systems. In A. P. Bishop, N. A. Van House & B. P. Buttenfield (Eds.), *Digital Library Use*. Cambridge, Mass.: The MIT Press.