REPORT OF THE RESEARCH INFORMATION MANAGEMENT (RIM) SCOPING COMMITTEE

June 2018

Michael O'Neal (Co-Chair) College of Earth Ocean and Environment
Monica McCormick (Co-Chair), Library
Jeffrey Caplan, Core Facilities
Dave Edwards, College of Health Sciences
Joe Kempista, Information Technologies
Aletia Morgan, Information Technologies
Cordell Overby, Research Office
Carl Schmidt, College of Agriculture and Natural Resources

Ex officio
Sharon Pitt, Information Technology
Anshuman Razdan, Research Office
Maria Palazuelos, Research Office

Background

The University of Delaware Research Information Management (RIM) Scoping Committee was established to frame an institutional perspective on research data management needs at UD, including the collection, management, delivery, and archiving of research data. The committee's specific charge was to assess the current data management climate at UD and to help develop the scope for future RIM-focused committees that will investigate specific challenges at greater depth, identify service gaps and priorities, and recommend solutions. Based on this charge, the Committee met three times during Spring 2018 semester and completed the following:

- The development and delivery of an online survey of faculty and researchers' current data development and management strategies, and the identification of unmet needs.
- An assessment of current UD storage solutions (see Appendix A)
- A thematic outline of resource/service needs for future committees to consider

Faculty Data Survey

The online survey we developed was delivered to all UD faculty and researchers (April 5th through the 30th, 2018). It was designed to collect information about current data management practices, resource needs and gaps, and knowledge of services (the survey questions are available in the link of Appendix B). We received 109 responses from faculty representing all of UD's colleges. The results capture the wide-ranging scale and complexity of data collected by UD researchers, as well as how they see their data management needs through the life cycle of a research project and beyond. A statistical summary of the online survey results is presented in Appendix B along with the committees' synopsis of the survey results. The key outcomes are clear in that a large number of respondents that suggested the following:

- data management is crucial to their professional advancement
- they welcome support in developing and implementing data-management plans
- they are unsure what support and resources UD is able offer
- there are great concerns regarding data sharing, backups, and security

Appendix C presents additional comments and suggestions regarding the online survey results that were developed as part of the discussion during the scoping committees' second meeting. We concluded that a variety of resources and services are likely needed to address the key concerns raised by the survey participants (i.e., data management planning; organization; backup and storage; varying levels of access and security; curation and metadata provision; long-term preservation). Outreach and training regarding UD's available services, for those whose projects allow them to take advantage of such resources, are also needed because researchers are not always well informed regarding current UD offerings, are unable to make effective use of them, or are not aware of data delivery or archiving requirements. However, it is also clear that any future campus-wide data management strategy requires a deeper understanding of the needs of active researchers in terms of data management policy and services to ensure we are planning for and provisioning resources in ways that meets the University's strategic goals.

Recommendations

We suggest that the Research Deans endorse the continuation of a university-wide RIM Services Committee that will further investigate faculty and researcher needs in a range of areas. Specifically, the committee suggests the following:

- Establish subcommittees to engage with researchers and support staff with expertise in specific topic areas (especially in terms of the outcomes identified in the online survey: plan development, backup, compliance, metadata, delivery, and archiving).
- Identify resource and service gaps and set priorities for addressing them in terms of what needs will be met by the University vs. those best handled at the departmental or college level.
- Incorporate (rather than duplicate) any ongoing UD efforts such as the those focused on high-performance computing and data science in addition to identifying research data management resources and service offerings across the university.
- Develop an improved outreach strategy in terms of directing researchers to resources and services at their point of need.
- Evaluate which data management needs are best served in terms of guidelines rather than regulatory compliance (i.e., incentives and resources that ease the path for ensuring knowledge of standards and ever-shifting regulations)

The committee members are in agreement that an improved data development and management strategy is necessary to raise the visibility of UD research and increase the likelihood of obtaining increased research funding. We do recognize that UD has a diverse array of support options offered centrally and within departments, colleges, centers, institutes and

core facilities, but also acknowledge we lack coordination and guidelines. Additionally, there are a growing and changing set of requirements from funders, including federal agencies, for data management planning, data delivery and access, and long-term archiving. Tracking these requirements and ensuring that they are met is a serious challenge. Appropriate solutions must emerge from a continued understanding of the problems our researchers face.

Appendix D offers a thematic outline of specific topics we discussed during our final meeting and we offer it to subsequent committees for expansion and/or prioritization as part of their evaluation and solution process.

Populating Future Committees

The RIM Scoping Committee members, including the co-chairs, are willing to serve on the follow-on RIM Services Committee. However, we also suggests that the Deans personally identify representatives from their own units, research institutes, and core facilities to participate on future committees to ensure a diverse and appropriate group that is able to offer insights to the thematic topics in Appendix D. Based on the experiences of the current committee, it is our recommendation that both faculty and technical specialists, who may be college IT staff, lab data managers, research faculty and/or graduate students comprise the subsequent committee. The co-chairs O'Neal and McCormick have volunteered to work throughout the summer with the Research Deans and other administrators to populate a list of potential committee members and return it for evaluation at the Research Deans Fall meeting.

Appendix A: Summary of Current UD Storage Services

This report was prepared by Anita Schwartz and Joe Kempista in March 2018:

https://docs.google.com/document/d/1OuywTZLlYIZcXY_YyrrBDSWbrijj81ua-GlTqWr8F8s/edit?usp=sharing

Appendix B: Online Data Survey and Results

Please use the following link to a PDF file with the questions used for the online survey and the subsequent Qulatrics results:

https://drive.google.com/open?id=1-BIpS_ur-9DItFEwxrm3w0HzYo-NFV-y

The following three sections summarize the responses of the web-based survey: Part 1 - General Information, Part 2 - Data Management Throughout the Research Life Cycle, and Part 3 - The Publication, Sharing, and Archiving of Research Data.

Section 1 – General Information (refer to Questions 1 to 9 in Appendix)

A total of 109 faculty and researchers from 31 different units completed the survey. This represents less than 10% of persons that were emailed the link. Despite the low numbers of participants, the diversity of colleges and units completing the form covers a substantial breadth of the University's diverse faculty and researchers.

The participants create data in diverse forms that are indicative of the breadth of research projects of the participants (i.e., code, text, databases, videos, images, etc.) (Q3). In terms of the volume of data produced (Q4), 23% suggested they were unaware of how much data they collected. However, 55% of those aware of the volume they produce suggested it was less than 100 GB per year.

We specifically identified those that use relational databases for their data storage and management (Q5) and found that 24% of participants relied on SQL Server, MySQL, Oracle, or another similar application for managing complex data structures. Given the specific nature of how relational databases are designed and managed, this outcome suggests there is reason to consider how UD would manage the long-term needs of archiving and/or sharing such information. N.B., the majority of participants using relational databases indicated that they rely on some version of Microsoft SQL server.

Questions 6-8 suggest that data is of clear importance to career development of faculty and researchers at UD. Collection, management, and dissemination of data is critical to tenure or promotion for 74% of participants (Q6). Another 74% of participants noted that their research funding or other obligations require a data management plan. However, 20% of participants suggest that they do not always operate with a well-defined data management plan (Q7), and 86% indicate that they do not receive help from their academic unit in the development of such plans (Q8).

Section 2 – Data Management Throughout the Research Lifecycle

Our request to identify how participants store and back up data indicates that 60% of the participants use cloud-based storage (Google Drive, Dropbox, Box.com, or Microsoft One Drive) as the primary means of storing and sharing their data (Q12). This simple solution logically

parallels the small amount of data collected by most researchers (Q4). Of the survey participants who are aware of current backup processes, manual versus automated approaches were split 50-50 (Q14). A subsequent committee should investigate whether manual back-up is from a lack of resources or by choice, and/or if there is a shortcoming in support for automated backup processes. Of those participants who backup data, 48% complete backups on a daily or weekly basis, while 21% of participants were unsure of their backup timing (Q15), highlighting concern that backup solutions remain unclear among those surveyed.

Given the significance of metadata, both in requirements for grants/contracts and in the long-term storage of datasets, our committee sought to specifically address participant awareness of metadata as well as in metadata creation. Of survey participants, 16% were unaware of what metadata is, while the remainder were evenly split between those who do or don't expect to produce metadata for their datasets (Q16). We as a committee are concerned that a bias against creating metadata, or a lack of knowledge of the significance of metadata, will impede archiving and dissemination of data.

Section 3 – The Publication, Sharing, and Archiving of Research Data

The focus of Section III was to assess data storage, management, and dissemination issues in the final stages of a project, in which archiving or future availability of data is a concern. Increasingly funding sources, project sponsors, and government agencies either request or require that project data be made publically available. However, half of survey participants note that their research data requires privacy (Q19) (i.e., human-subjects considerations or emerging patentable discoveries). At UD, most (73%) participants typically only allow constrained access to data, such as within a research group, after project completion (Q17). Most (55%) see the publication process as the primary pathway for public data dissemination, while smaller numbers share their data via archives, either open-source (11%) or university-level (5%). Q20 provides mixed responses regarding data delivery, suggesting that researcher requirements and values are complex. A key investigation for future committees would be to identify the personnel and resources that would be required on UD's part to facilitate both greater privacy support and greater dissemination efforts, in addition to managing timescales for data hosting.

When provided with a list of support and infrastructure needs, over 2/3 of all participants found ***each*** of needs listed to be important or critical for the University to offer (Q22). Participants were asked to specify data management needs in a free-form text (Q24). These responses range from concerns of vulnerability to general need for assistance across the breadth of data management and development issues. Many participants emphasize an urgent need for custom backup solutions, consulting, and support. A follow-up question to query self-assessed need for assistance (Q23) shows that the participants would make use of UD-provided support (in the form of workshops, consultations, and online resources) for a wide range of data management topics.

Appendix C: Committee Discussion/Suggestions Directly Related to the Online Survey

The RIM Scoping Committee recognizes that the faculty require help in many aspects of data management. However, it also recognizes that our services should attract faculty to better use UD resources rather than being offered in the framework of enforcement.

The strongest recommendation would be that subsequent committees follow up more information-gathering, including interviews with researchers who have specialized data needs to identify shortcomings or missing components of the data management life cycle that could be addressed institutionally.

Because the majority of responders to the survey recognized that data collection and management was critical to their promotion at UD, data management training as part of new faculty orientation would be particularly efficient and a great benefit.

Based on the specific comments offered by survey participants, backup is an obvious and primary concern of most faculty/researchers. Not only the backup itself, but the security of the backup. Given the propensity of faculty to utilize current cloud-based resources, it is worth knowing whether or not faculty are aware and/or satisfied with UD's relationship with Google and other cloud-based resources supported by UD.

Many faculty consider their data to be very private during the research cycle, or have contractual or legal obligations to keep data private and/or anonymous.

Many universities have physical worksheets for faculty and researchers to complete as they plan research projects that are data intensive. This type of activity, coupled with perhaps a series of ongoing data management workshops provided at the university level, might be a critical step in improving/simplifying the organizational framework throughout a project life cycle.

UD could easily consider the value of a website that would direct faculty to time-of-need solutions, modeling off well-designed web sites from peer institutions:
https://libcms.oit.duke.edu/data/data-index Duke University
https://finder.research.cornell.edu/storage Cornell
These sites provide an array of data management services and solutions (ranging from workshop schedules to available software and its use and implementations, tips for backup strategies, storage resources, etc.).

Appendix D: Topical Areas for Further Investigation

The following outline presents a list of thematic topics that the committee provided as part of the discussion during its final meeting. The topics serve as a starting point for the future committee in terms of general areas of interest and/or questions that require prioritization as the next committee assesses data development and management solutions.

Physical Infrastructure and Software Solutions

- Storage and backup
  - Varying requirements for networking and speed
  - Variations across campus locations
  - Size requirements, speed requirements
  - Servers provided at what level: central, college, individual lab/PIs
  - Different requirements for active vs archived data
  - Reliance on non-UD backup options (collaborators, discipline-related storage)
  - Google and other cloud services vs campus-managed
  - Automated backups
  - Scratch space

Compliance and Security

- Public access
  - Funder requirements
  - When / for what kinds of data is this required?
  - When is this restricted?

- Human subjects
  - Special needs for security
  - De-identifying data for public access

- Security and access
  - Variations across the research cycle
  - Ability to collaborate beyond UD during research
  - Public access requirements
  - Privacy requirements
  - Regulatory requirements (by industry, funders, etc.)

Service provision

- Tiers of service
  - Differences because of scale/size of data sets
  - Support at the lab, department, college, or campus level
  - Charges for amounts of data stored?

- Data management planning
  - Funder requirements
  - Training
  - Service implications for implementation of plans

- Provision of persistent identifiers
    - DOIs for data sets
    - Permanent links
    - ORCIDs
    - Integration into other services

- Metadata support
    - Required for effective use, preservation, and discovery/access
    - Gaps in knowledge
    - Varying needs at different points in the research cycle

- Long-term preservation
    - Backups and long-term preservation
    - Disciplinary or funder options external to UD
    - UD preservation options

Policy

- Institutional research data guidelines
    - Are there written guidelines and variations across campus?
    - What happens to data when faculty leave UD, or arrive from elsewhere?
    - Who is responsible for addressing guidelines and compliance?

- Awareness of regulatory requirements
    - How to ensure this is available to researchers
    - How to ensure this is up to date

Outreach and Training

- Awareness of service options
    - Effective online information that points to support and service choices
- Training options
    - Services for relevant points in the hiring and research cycle (information sessions, workshops, consultations)