Data Management and Data Curation Pilot Report and Recommendations

Prepared by Mary Bell and Christine Kollen

Executive Summary

From the fall 2015 to the spring 2017, the Research Computing Governance Committee (RCGC) Data Management and Curation Subcommittee conducted the Data Management and Data Curation Pilot in collaboration with the UA Libraries (UAL), University Information Technology Services (UITS), and the UA Office of Research, Discovery, and Innovation (RDI). The pilot worked with five research projects through the data life cycle to evaluate data management services. This report and recommendations is based on the result of the pilot, what worked well, what did not work well, training needs, and feasibility of implementing these services campus-wide. The funding for this project was provided by the RCGC sponsors and included: data management and curation staffing support, purchase of servers and storage (to develop a pilot instance of iRODS), and staff support for the servers and storage. The project staff reviewed and made recommendations on each project's data management plan (DMP); developed templates, workflows etc.; set up data management technology and tools; provided metadata support; and worked with the research projects to help them prepare their data and software (if applicable) for deposit in a repository.

By developing an efficient set of data management and curation services that include training and support for researchers and the needed technology infrastructure for these services will allow researchers to devote more time to research. The services will also support the goal of developing a research data ecosystem that facilitates data reproducibility, and have a positive impact on the UA's overall prestige and success in obtaining grant funds.

This report includes an environmental scan, an overall summary of the pilot, details from each project, results of exit interviews with the project participants, and recommendations.

General recommendations include:

- Support needs to be a joint effort between the Libraries, UITS, and RDI.
- Data management and curation services need to be fully developed, robust, agile, and available at point of need.
- Researchers feel overwhelmed and want to devote as little mental bandwidth or executive function as possible to the tasks and tools surrounding data management. Ideally, it should be as transparent as possible to the researcher, and integrated into existing workflows.
- Researchers desire customized support that works with tools the researcher is already familiar with (as much as possible).
- Concrete solutions are better: researchers wanted checklists, rubrics, templates, protocols, and tools customized for their needs.
- Data management best practices are especially crucial the larger the research group.
- Most research on campus is not "big data" research, and may not need HPC resources. However, they need secure working storage and a long-term archival solution.
- For UA Health Sciences researchers, HIPAA requirements make data management and disposition more complex; they may be more willing to accept a fee-for-service model than main campus researchers.

Specific recommendations include:

- A suite of services should be offered at point of need during different points in the research lifecycle.

- Offer a service like Open Science Framework (OSF) for Institutions, and encourage use by offering 20 hours of data management consultations to grant startups that agree to use OSF as their shared data storage and communication system.
- Coordinate services with the Office of Sponsored Projects and other units so that the library is notified when grant proposals are submitted/and or funded, or at various project review points, which could automatically trigger an email to researchers offering data management services and resources.
- Investigate campus partnerships with UITS, RDI, and CyVerse in developing a data repository.
- Develop a campus data repository – provide management, sharing, and preservation of research data in a structured infrastructure.
- Investigate electronic lab notebooks to use as a collaboration platform. Conduct survey of what is currently being used, what needs do researchers have. Is there an ELN that the campus should support?
- Add staffing, training and funding for liaisons librarians, who could suggest discipline-specific resources to help with data discovery and archiving, and conduct departmental data management workshops and consultations.
- Continue and expand training opportunities for graduate students and postdocs to acquire skills in data management and curation in their discipline. Consider developing an online for-credit course and work with instructors (in conjunction with liaisons) to integrate data management into discipline-specific courses at the graduate level.
- Develop online data management tutorial modules.
- Update and expand the UA Data Management Resources website
- As the Data Curation Network (DCN) is implemented (see page 6), consider how their data curation services could supplement and enhance the services we offer at the UA.

Introduction

In 2011, the University of Arizona Libraries and the Office of the Vice-President for Research (VPR) appointed the Campus Data Management and Curation Advisory Committee to address the need for a University-wide strategy for data management in support of research activities, to identify needed support mechanisms and cyber-infrastructure, and to propose organizational models and distribution of responsibility. The Committee's recommendations (approved on April 6, 2012) included: develop a coordinated, campus-wide strategy in support of data management for researchers across campus with the UA Libraries (UAL) serving as integrated point of service for faculty and researchers; University Information Technology Services (UITS), in collaboration with the UAL and the VPR's Office, should work to devise an overall long-term strategy for the support of data storage, data access, and data preservation in support of UA's research needs; and establish an ongoing Campus Data Management Subcommittee of the Research Computing Governance Committee (RCGC). The full document, "Recommendations of the Campus Data Management and Curation Advisory Committee" is available at http://data.library.arizona.edu/sites/default/files/final-recommendations_0.pdf).

One of the accomplishments of the RCGC Data Management Subcommittee was to conduct the Research Data Management Survey. The survey was a collaboration between RCGC and the UAL and distributed during the spring semester 2014 to faculty, researchers, graduate students, post-doctoral fellows and others. The purpose of the survey was to discover how research data is being managed across various units at the university, determine what the demand is for existing services and identify new services that UA researchers need. The survey asked questions related to the participant's demographics, data storage, data management, data sharing and data reuse. The report, "Research Data Management Survey Report and Recommendations" is available at https://arizona.box.com/s/lnmplngv34ofjfgykp8i8rzpb4ddq440.

In order to explore these needs in more detail, the RCGC Data Management and Curation Subcommittee proposed and received funding from RCGC sponsors to conduct the Campus Data Management and Data Curation Pilot. (The proposal, "Data Management and Data Publication and Curation Pilot" is available in the Appendix.) The pilot worked with five research projects, from fall 2015 to the spring 2017, through the data lifecycle to evaluate the implementation of data management services. This report and recommendation is based on the results of the pilot, what data management services were needed by researchers involved in the pilot, what worked well and what didn't work well, training needs, and the feasibility of implementing these data management services campus-wide.

Environmental Scan

Research data has value beyond the original purpose, it promotes innovation and new data uses, ensures reproducibility and validates results, and fosters interdisciplinary research. This potential oftentimes falls short of the promise. Research data may lack documentation and metadata, suffer from digital deterioration, and not make it beyond the researcher's domain to a wider audience. The scholarly community values well-curated data, it makes it easier for others in the field to understand the data, it is more likely to be trusted, and be reproducible. As a result, requirements for how to effectively manage, share, and preserve research data have emerged. Researchers are under more pressure from funding agency mandates, their institutions and from their disciplines to make their data findable, accessible, interoperable, and reusable or FAIR (Wilkinson et al., 2016). There has been an increase in the U.S. in the number of federal funding agencies and some private foundations that require researchers to provide public access to the results of their research. In 2013, the White House Office of Science and Technology Policy (OSTP) released a policy memorandum, "Increasing Access to the Results of Federally Funded Scientific Research," directing federal agencies with more than $100 million/year in research and development expenditures to develop plans to make publications resulting from that funding publicly available. The memo also called for funded researchers to better account for and manage the digital data coming from their funded research. As of 2016, nineteen agencies have begun implementing public access plans. The majority of these agencies require that researchers include a data management plan (DMP) as part of their grant proposal and to openly share their data in a "public data repository", either disciplinary or institutional, at the end of the project. In addition, there is an increase in the number of journal publishers requiring authors to submit their supporting research data along with their manuscript as part of the peer review process. Some examples include PLOS One (http://journals.plos.org/plosone/s/data-availability) and the Nature Publishing Group (http://www.nature.com/authors/policies/availability.html#data).

Research Data Management Services

In the Association of Research Libraries (ARL) *SPEC Kit 334: Research Data Management Services*, Fearon, Gunia, Lake, Pralle, and Sallans (2013) reported on a survey of ARL member libraries about research data management (RDM) services and identified two emerging areas: research data management and data archiving. Seventy-four percent (or 54) of the libraries who responded to the survey offered research data management services, while another twenty-three percent were in the planning stages (13).

Of those 54 libraries currently offering RDM services, the following are a breakdown of specific services:

RDM Services (N=54)	**N**	**%**
Online DMP resources	47	87%
DMP training	33	61%
DMP consulting	48	89%
RDMS besides DMP support	53	98%
Data archiving by library	40	74%
Data-specific archive (other than institutional repositories)	5	9%

Of the libraries responding to the staffing question (N=53), 51% provide RDM services by a committee of staff from departments within the library. Other staffing solutions include a campus-wide committee that include staff from the library (9 or 17%), a single position within the library (8 or 15%), and a department within the library (6, or 11%). The size of these committees or groups vary from 6-10 staff depending on where the staff come from (campus-wide, library, or one library department). Eight of the respondents indicated that they have only one position that provides RDM services. There are only a few of these positions that have RDM services as their only, or even their main responsibility (17-18).

The vast majority (98%) of libraries indicate that they fund RDM services from their internal library budget. From 2-11% of respondents indicate that they also receive some funding from other sources, such as administrative funding, external grants, temporary or special project funds, research project funds, etc. Only 4% of institutions reported that they charge a fee to a researcher or researcher's grant for RDM services (77).

Data Curation and Data Repository Services

In *If You Build It, Will They Fund? Making Research Data Management Sustainable*, Erway and Rinehart (2016, 5) note that funders are beginning to look to institutions to provide solutions. They prefer preservation solutions provided by an institution's library, as compared to the access nodes that are being provided by the majority of disciplinary data repositories. However, in the ARL *SPEC Kit 354: Data Curation*, Hudson-Vitale, Imker, Johnston, Carlson, Kozlowski, Olendorf, and Stewart (2017, 4) state, "library technical and human infrastructure are just now reaching the point of accepting and curating data". Providing data curation services has not translated yet into strong staff levels, with most institutions placing responsibility on individuals with other duties to carry out. Of the 80 ARL libraries that responded to the survey, the "majority of ARL libraries are providing data curation services or that development of these services is underway". Of those 51 institutions, 90% also provide repository services for data, using their institutional repository or have developed a stand-alone data repository. Repository platforms vary, with 22 of the reporting institutions using DSpace, 11 using Dataverse, 10 using Fedora/Hydra, 7 using Islandora, and 17 using a combination or other platforms (3).

In *SPEC Kit 334*, Fearon, Gunia, Lake, Pralle, and Sallans note, "most of the libraries with data archiving services (84%) are absorbing those costs through their internal budgets". 84% provided data archiving services out of the library's internal budget, 24% funded data archiving through grants, 14% charge researchers, and 19% found other funds" (15).

In *Research Data Stewardship at UNC: Recommendations for Scholarly Practice and Leadership,* the Provost's Task Force on the Stewardship of Digital Research Data (2010, 62) reported on a comprehensive survey of faculty in 2010 where they asked, "In your opinion, where should the funding come from to cover the costs of data management and storage for research supported by grants, contracts, or other external sources of funding?" Over half of the 2010 respondents answered, "It should be paid for by the University from overhead/F&A funds it receives from grants and contracts." In addition, when asked where funding should come from for research not funded by external sources, 63% thought it should come from university funds.

A recent project, *Data Curation Network: A Cross-Institutional Staffing Model for Curating Research Data* (Johnston, Carlson, Hudson-Vitale, Imker, Kozlowski, Olendorf, and Stewart 2017, 8), propose a solution to the challenge of lack of local expertise in discipline-specific data curation. "The curation staff, or the 'human layer' in the repository stack, bring the disciplinary knowledge and software expertise necessary for reviewing and curating data deposits to ensure that the data are reusable. … it is unrealistic to expect that every academic library can hire a data curator for every data type (e.g., GIS, tabular spreadsheets, statistical survey, video and audio, computer code) or discipline-specific data set (e.g., genomic sequence, chemical spectra, biological image) an IR might encounter." The Data Curation Network (DCN) addresses the challenge of providing subject-specific data curation services collaboratively across a network of multiple institutions and digital repositories beyond what an individual institution alone might offer. The DCN will help libraries provide critical new or more comprehensive data curation services, participate in development of shared standards and practices, and stabilize services during staff transition or shortages.

The ARL *SPEC Kit 334* (Fearon, Gunia, Lake, Pralle, and Sallans 2013, 81-82) reported the following top five challenges in delivering RDM services:

- Campus collaboration: Libraries indicated struggles with getting buy-in from various campus constituencies, and that silos within the university make collaboration difficult. This included the separation of IT services and Research and Development services from library services.
- Limited Staffing and Funding: Libraries uniformly report lack of staffing as a problem, since most libraries have asked staff to add RDM services on top of their other full-time duties, rather than adding staff. As well, RDM services require data skills or subject area knowledge outside the expertise of some liaison librarians, so training costs and time must be added as well. This lack of staff resources and funding significantly impacted libraries' ability to deliver RDM services, and their ability to sustain the services they currently offer.
- Faculty engagement: Libraries reported difficulty with getting faculty to engage with RDM services, due to faculty schedules, lack of buy-in, lack of interest, and reluctance to share data.
- Infrastructure: Comments mainly relate to data storage services and support for both working storage and long-term archival storage.

Conclusion

As the ecosystem of research data management and curation continues to evolve and become more complicated, it will be critical for academic libraries to enhance services that support data management and curation. How will libraries' staffing models continue to evolve, will more staff be devoted to these activities? What effect will the Data Curation Network and similar initiatives have on services libraries provide? What other funding models besides internal funding can libraries explore. If grant or other temporary funding is used (overhead or facilities and administrative costs), how should libraries address the sustainability of these funding sources? Universities may need to look at a combination of these funds in order to expand data management and curation services.

Data Management and Data Curation (DMDC) Pilot Project

The Data Management and Data Curation Pilot worked with five research projects, from fall 2015 to the spring 2017, through the data lifecycle to evaluate the implementation of data management services. The funding supported data management and curation staffing support, purchase of servers and storage (to develop a pilot instance of iRODS), and staff support for the servers and storage. The Libraries, CyVerse, and UITS worked together on integrating the various data management and curation services and tools provided to pilot participants.

Table 1: DMDC Pilot projected milestones, status updates

Subactions	Timeframe	Metrics	Status updates
Appoint DMDC Pilot planning team	Fall 2015	Members appointed	Team in place - November 2015
Planning team sets up application process and selects pilot participants	November 2015-February 2016	Three-six projects selected	Six research projects selected for the pilot
Interview pilot participants	Spring semester 2016	Interviews conducted and transcribed	Interviews were completed in March 2016
Develop final recommendations for each participant	By end of Spring semester	Recommendations finalized	May 2016
Evaluate recommendations for each project and make any needed changes at midway point	November 2016	Verify utility of recommendations	Throughout the project
Work with researchers to get their data ‘repository ready’	TBD	As appropriate, data is uploaded to data repositories	At various times
Evaluate pilot and provide recommendations	January 30, 2017	Report on results of the pilot and feasibility of ramping up services to campus	March-June 2017

Overall Pilot Timeline

The library hired a Postdoctoral Research Associate to serve as the embedded data management librarian (data management and curation staffing support) for the pilot’s research projects. Along with the Data Curation Librarian, the postdoc’s role included reviewing and making recommendations on each project’s data management plan (DMP); developing templates, workflows etc.; setting up data management technology and tools; providing metadata support; and working with the research projects to help them prepare their data and software (if applicable) for deposit in a repository.

Pilot Planning Committee (David Moore, Kim Patten, Susan Miller, and Kathleen Bowles) and pilot project team (Chris Kollen and Jodi Reeves-Flores) recruited campus research projects through the UA’s Competition Space. Only six projects met the pilot criteria (see Pilot Notice Application at https://arizona.box.com/s/mkywu06tukktzmohflr32bnn8umtjuy3), the selection committee chose all six to participate.

The project team interviewed (and recorded) each of the six project teams to get information about their data workflows and management needs. The questions were customized for each project based on their original application. A student worker then transcribed the interviews.

From the interviews, the project team compiled for each project a list of deliverables (data management needs) that the participants discussed, edited, and agreed to. The finalized recommendations and deliverables were documented in May 2016.

At the beginning of the pilot, the project team contacted the UA Human Subjects Protection Program, which determined that the project would not need to submit an IRB application. In the application for participating in the pilot, we specified that any published work resulting from the project would only refer to pilot participant projects in general terms. As a result, the projects are described only in general or in aggregated terms (see Table 2).

Subsequently, the pilot staff met periodically with the project participants to implement each project's deliverables. This turned out to be much more involved than anticipated, especially for projects in which we were meeting with the PIs rather than research staff. Scheduling meetings was often difficult, and sometimes resulted in significant delays, as many of the project participants were out in the field for weeks at a time, had grant or tenure deadlines or travel that made them unavailable. One project participant dropped out in April 2016 before we started to work on the project's deliverables. Additionally, Jodi Reeves-Flores (the Postdoctoral Research Associate) left at the end of June 2016 and was replaced by Mary Bell at the end of September 2016. That set the project timeline back a few months. However, all of the projects were completed by April 2017. The results were analyzed, and the report and recommendations written.

Project Descriptions

The six projects selected for the pilot ranged from very small (2 persons with ~250 MB of data), to large and complex (40+ collaborators across multiple labs and locations with ~20 GB of data), to very large with heritage data (longitudinal data over 25+ years, ~ 5 TB of data). Four projects had collaborators at other institutions. Five projects had sensitive or protected data. One project was fast-tracked, while others had been going for decades and had legacy data. Two required extensive fieldwork. All had multiple data types and data workflows. Almost all struggled with document control and data workflows, especially when students were involved. Only one project was at the beginning stages when it was selected (Project E), but by the time we were able to get the pilot started it had already moved to data collection, and we had trouble keeping up with the pace of the project. Finally, one project withdrew from the pilot (Project F) after we interviewed them and developed recommendations but before we started working with them on the deliverables.

Table 2: Data Pilot Projects Overview

Project and general field	Primary contact role(s)	Project Personnel	Data types	Estimated data storage needs	Length of project	Sensitive/ protected data?
Project A Entomology	PI & lab manager	1 PI, 3 co-PIs (at other universities), lab manager, and 5+ collaborators (postdocs, graduate students, undergraduates)	Proprietary lab equipment formats, Excel, images	40 GB	5 years	no
Project B Engineering Education	PI & graduate student	1 PI and 1 collaborator (grad student)	Excel, CSV, python scripts	250 MB	3+ years	Yes, IRB, FERPA
Project C Wildlife Biology	Research Scientist	1 PI, 2 research scientists, 2 student workers, and 5+ collaborators (postdocs, grad students, undergrads)	Access, shapefiles, ArcInfo coverages, images	~5 TB	25+ years	Yes, protected (endangered species)
Project D Cancer Research	Data Manager & Research Admin (3 FTE)	4 PIs, 5+ collaborators at three other institutions	images, murine models: Access, Excel, SAS, Stata, SPSS	~50 GB	30+ years	Yes, HIPAA
Project E Public Health	Data manager, PIs	2 PIs, 1 data manager, 4 co-investigators, 40+ students & community volunteers, and a research group at another university	All research data in RedCap	~20 GB	2 years max	Yes, IRB (standard and tribal)
Project F* Education	PI	1 PI, post doc, 5+ collaborators	Access, Excel, images	~300 MB	3 years	Yes, IRB (standard and tribal)

*Project F withdrew from the Pilot. They had already finished data collection when we interviewed them.

Project Deliverables Summary

Deliverables (data management needs) ranged across most of the data management lifecycle, with a significant portion (51%) related to data management and organization issues, another 26% related to open access data publishing support, and 10% to questions about data archiving. The other 13% were concerned with technical skills such as R, Python, and other analysis tools; with study closure procedures and checklists; and with a request to review a data management plan for a grant proposal (see Figure 1).

Figure 1: Percentage of deliverables by topic

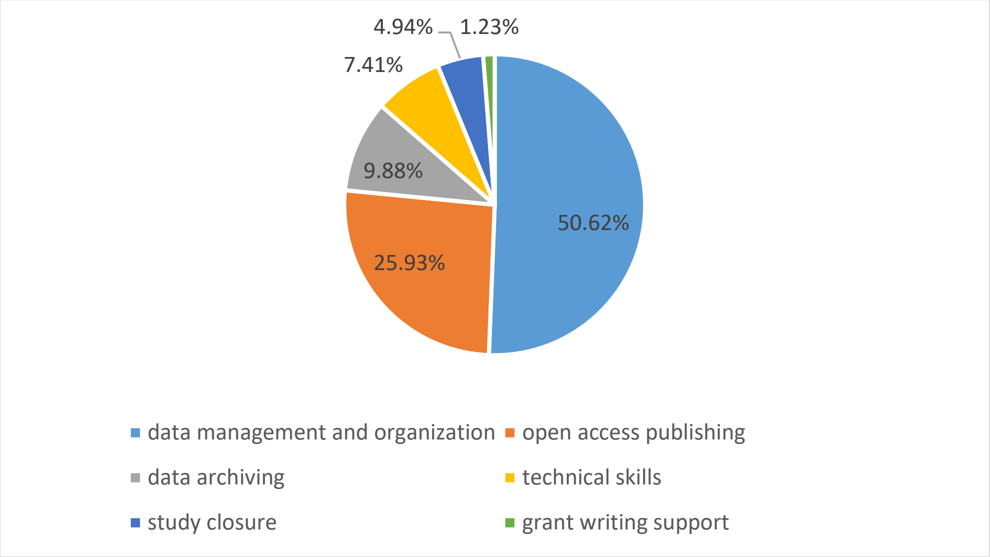
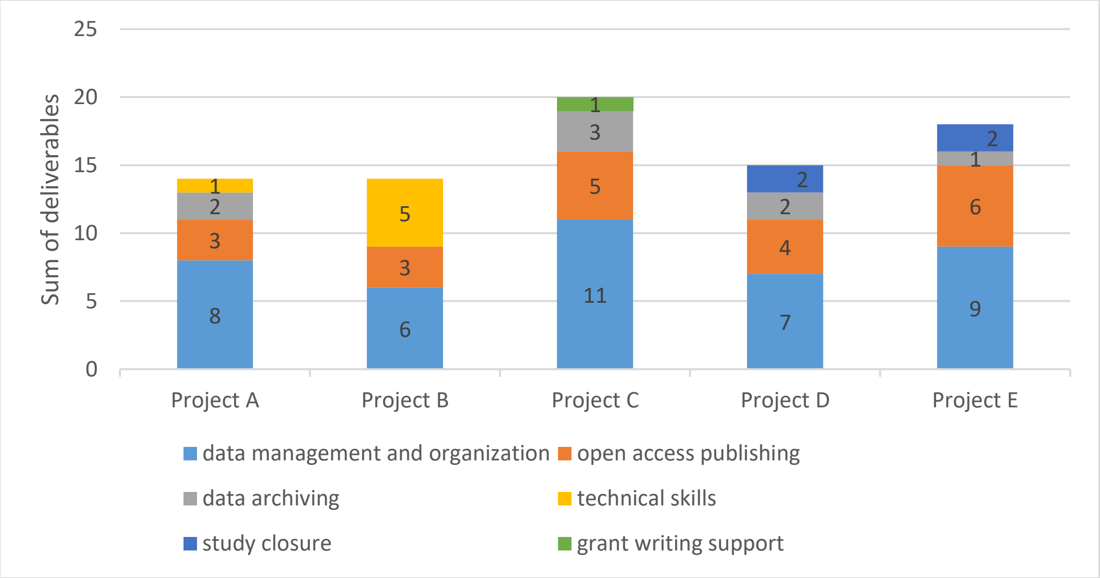


Figure 2: Number of deliverables by topic and project



Project Narratives

Project A

Project A's disciplinary area is in entomology with one PI, three co-PIs (at other universities), one lab manager, and 5+ collaborators (postdocs, graduate and undergraduate students). This project has been running for 5 years with approximately 40 GB of data. Data is collected from five pieces of equipment in the lab. The data is transferred and stored on a central computer.

Pre- pilot – status of the data

There were a multitude of ways that post-doctoral research associates, graduate students, and undergraduates were working with the PI. The post-doctoral research associates and the Ph.D. students were sharing their data with the PI using DVDs; at times the DVDs would get lost. They have data on data sheets or lab notebooks that are then entered into Excel. The PI said that he is losing control of his data. For some data, the PI did not have access to it, a previous lab manager had set up individual folders for students working in the lab. Once the students left, the PI no longer had access to the files. The students wrote down values in a lab notebook for three pieces of lab equipment; the values were later entered into excel.

The PI backs up his data to Dropbox (shares with co-PIs at three other institutions). It is not clear if he makes any other back-ups.

Identified problems with Data Management

The PI identified issues around how post-docs, graduate and undergraduate students in his lab collect, label and transfer data to the PI. How is data collected by the lab equipment transferred to the central lab computer, and how are the data backed up?

Pilot Project A Working Recommendations

The DMDC project team will provide guidance and support to improve research data management practices for the PI's lab. The project team will document research data management workflows that document how students should collect, label and transfer their data to the PI. This will include protocols for how students transfer analyses and data when they leave the lab. The recommendations will also include documenting file naming, versioning and organizational systems, building on current approaches to make sure good practice is followed. The project team may also suggest exploring the development of an Access database that could take in all of the datasets. As the DMDC Pilot progresses, we can also explore support for archiving finished datasets. The participant will agree to meet (bi-weekly) to workshop each of the deliverables below until the developed protocols are in place and running. The PI will also agree to provide feedback on how the protocols are going.

Project A – Pilot Implementation timeline by deliverable

Deliverable 1: Set up iRODS for storing and managing data
The PI was originally interested in using iRODS as a back-up for his data. The project team worked with Project C first who was also interested in implementing iRODS. The PI later decided that iRODS was overkill for his project and decided not to pilot it in his lab.

Deliverable 2: Document data management workflow
The project team will establish workflows that document how students should collect, label and transfer their data to the PI.

June 30, 2016 – Post-doctoral research associate leaves.

July 2016 – PI out for most of July, schedule meetings with PI and his lab manager starting in August 2016.

September 2016 – developed diagram of data workflow showing each piece of lab equipment and that the data is saved to a central lab computer. Also included recommendations on file naming, version control, and data documentation.

There are three pieces of lab equipment where the students write down the reading in a lab notebook then enter it into excel. The project team recommended that whenever possible the PI set these up so that the data is collected by the equipment that then can be exported instead of running the risk of transcription errors. The scale is the one exception - the students record weights from the scale in a paper notebook. The scale is so delicate that the PI decided that it is better to write down the value first before inputting the values into excel.

Deliverable 3: As the pilot progresses, the project team can also explore support for archiving finished datasets.
The project team would also establish a method for archiving datasets that are no longer being added to.

The PI worked with the Director of the Campus Repository Services to add an excel file associated with a journal article.

Project A - Exit Survey Results

The project team was unable to schedule an exit interview with the PI.

Project B

Project B is a project in online learning materials for a chemical engineering class with one PI and one graduate assistant (GA). This project has been running for 3 years, with funding from a couple of different sources.

Pre- pilot – status of the data

The data are in approximately 100 excel files (4.2 million data points); however, they have not yet been anonymized. The data is exported from D2L using a python script that was developed by the PI’s graduate assistant. They back up their data in Dropbox.

Identified problems with Data Management

They are interested in moving the data currently in excel into a SQL database. They need to document the python script so another graduate assistant could come in and run it again (for subsequent semesters). The data also needs to be de-identified and they need help with file naming and versioning.

Pilot Project B Working Recommendations

The project team will work with the PI and GA to document the research data management workflow, which will need to cover how the python script works, where data and derivatives are stored, plans to de-identify data and file naming and versioning conventions. The project team can also provide guidelines on merging data, if needed.

Project B – Pilot Implementation timeline by deliverable

Deliverable 1: Document Research Data Management Workflow

- June-July 2016 – reviewed IRB and provided comments to the PI.
- Using Box to store their data will meet requirements for confidential data as outlined by UITS (http://security.arizona.edu/data-classification-and-handling-standard).
- June 2016-July 2016 – provided feedback to PI and GA on documentation of the Python script. GA developed two documentation files. The first file, DMDC Pilot Plant Program Roadmap, was used to log ideas, features, and plan for the Student Data Mining Program. Each is coded with Completed, Working On, and have not started/will not implement. The second file, Pilot Project Read Me File, provides documentation of the Python script.

Deliverable 2: Facilitate SQL support if this is a priority

- May 2016 – PI decided that SQL is no longer a priority for them, they will keep the data in excel and use R to process/analyze.
- June 2016 -- GA attended the Software Carpentry workshop (organized by CyVerse) and found it helpful with R, Python, and networking in general.

Deliverable 3: GA will contact Tech Transfer regarding patenting or licensing the python script.

- GA contacted Tech Transfer, not sure of the outcome.

Deliverable 4: Get Information on visualization and analysis support

- Provided information to the PI and GA about UITS Research Computing's visualization consulting services. Also recommended that if needed, they contact SBSRI for assistance.

Deliverable 5: Make suggestions for publishing data and provide information on data sharing requirements

- June 2016 – developed a list of potential journals for the PI to publish his research study in. Used the Journal Citation Reports on the web – Science and Social Sciences edition on 6/2/2016. We recommended two journals: Review of Educational Research (Journal Impact Factor = 3.897 and Total Cites = 4,649) and Journal of Research in Science Teaching (Journal Impact Factor = 3.162 and Total Cites = 4,410).

Project B—Exit Survey Results

The project team was unable to schedule an exit interview with the PI.

Project C

Project C is a project in wildlife biology with one PI, two research scientists, two student workers, and 5+ collaborators (students and postdocs). This project has been running for 25+ years, so there is a large amount (~5TB) of legacy data, mainly in older versions of Access and Excel, along with geospatial data and image data. Some of the data is sensitive since it deals with endangered species, so there are limits on what kind of data can be shared in open access repositories.

Pre-pilot – status of the data

The data included in this research project consists mainly of telemetry data, collected in the field and entered into Access databases on a field computer. After various quality checks, the data is exported to Excel and put on the lab computer, where one of the scientists imports it into an analysis program and exports the data as GIS shapefiles. The PI and various graduate students use the data for research studies.

The lab was in the process of moving buildings when the initial pilot interview was held, so the data backup and storage plan was in flux. Previously the data set up was a LAN with five desktops: one each for the PI and the two scientists, and two student computers. “So we had that LAN and we had the 2 student computers were like psuedo-servers. One computer had all the project data, one computer had all the GIS data and we would just map those drives to the network . . . we would always map it as the same letter drive all the way around and same with all of our Access databases so ArcMap can always find G://....no matter what computer we're looking at. … I was backing up everything to a NAS (network attached storage) that had RAID.”

During the move, the Access, Excel and shapefiles were backed up to external hard drives every night from the lab computers using third party backup software, Easeus.

Once the move was complete, the system was moved to two 16-TB NAS boxes in a climate-controlled room: one holds the data, the other one holds the backup. They are trying to “set it up as a proper server with permissions,” since the group has grown beyond the original 3-4 people.

They have not used cloud backup, because some of the data is sensitive. They had tried Box when the University first introduced it, and found it too slow to upload files. They thought about using Dropbox instead, if they were working with someone in another location, but they are aware Dropbox is not secure. They are open to revisiting Box as a backup solution.

Older data is sitting in an untouched Access database files or Excel files, along with the active datasets. Some are in Access 97, some are in Access 2003; the goal is eventually to migrate everything to Access 2010.

Identified Problems with Data Management

1) Versioning of files: example of a powerpoint that did not contain updated information, which was in an email but not on the hard drive; consequently they grabbed the wrong version for a presentation.

2) Lack of file access/management protocols: They have multiple graduate students accessing data; the PI pretty much does things haphazardly so there are things on his laptop that do not make it onto the system, etc. They keep the data on the server, but everyone has their own versions on their personal computer for presentations, analyses, etc. that are not updated on the main server.

3) Managing graduate student datasets and documentation when they leave.

4) File sharing: Just signed a letter of understanding with another university, and they will be sharing and exchanging data with them. They may need a cloud storage/file sharing solution in the future.

5) Migrating/archiving/sharing old data

Pilot Project C Working Recommendations

The project team will provide guidance and support to improve research data management practices for Project C. They will: provide a management workflow for the telemetry data collected; develop an iRODS system for the project to help with managing data (tentative); develop templates for archiving graduate student data; and provide support in archiving datasets no longer being added to. The project participant (primary contact) will agree to meet (bi-weekly) to workshop each of the deliverables below until the developed protocols are in place and running. They will also agree to provide feedback on how the protocols are going.

Project C -- Pilot Implementation timeline by deliverable

Deliverable 1: The project team will liaison on the project's behalf to explore implementing iRODS in order to establish one place where lab members can save and access data, with the ability to control access.

Setting up the iRODS server: the process took considerable time (9 months), and by the end of the pilot, it was set up as a backup system, but project participant did not see the value in it.

- May 2016--Met with project participant, she decided to try using iRODS with a current project. Project team met with UITS. They will set it up so that versioning uploaded files is automatic, and files are locked for editing. A better version of the interface will be available in a couple of months. The project team wrote up instructions on how to set up an account and access iRODS and sent it to project participant.
- June 2016—UITS set up iRODS: not able to do file locking, but versioning set up. Project C participant uploaded some files. Delay while UITS person was unavailable. Project participant tried out iRODS, but the current interface did not have enough functionality to implement for their project. Issues include not being able to lock a file for editing and not being able to edit a document without first downloading it then uploading the new version. Metadata fields were a plus. Might be useful for sharing or storing data that is not in the process of being modified. Can revisit once the new interface is in place.
- June 30, 2016 - Postdoc leaves
- August 2016—Discussed possibility of having project students scan their notes and add to iRODS. Decided a better solution would be to set up a data scan folder on the LAN: have students deposit scanned field notes there, then project C participant will upload to iRODS to use as a backup. Decided to wait to use iRODS until UITS has completed updates for the discovery layer.
- September 2016 - new postdoc starts
- October 2016—Discovery environment front end still not functional. Will need a programmer to set up rule development. UITS will upgrade iRODS. Send pilot staff information on WebDAV and Cyberduck interfaces, and basic iRODS information. Sent iRODS information to project C participant.
- January 2017—experimenting with Cyberduck, got it working, sent it to Project C participant to use.
- February 2017---Project C participant finally got updated iRODS installed and working: will use for backing up large static files, i.e., telemetry files, including photos.

Deliverable Two: data management workflow; the project team, with input from project researchers, will focus on creating a data management workflow for the telemetry data currently being collected.

- May 2016 – Project C participant was going into the field, would take notes on the data collection process.
- June 2016 -- Project C participant sent us a selection of current documentation for the data. Reviewed and sent feedback. Data collection protocols need to be consolidated so it is easier to access and update them. Set up protocol to scan field notes. Set up data scan folder on the LAN for students to deposit scans. Document scanning procedure in a README file. Helped create documents for data collection and processing workflows.
- Explore metadata standards: sent Project C participant the Dublin Core standards used in the Campus Repository. Also conferred with Metadata Librarian, who suggested the free program

Morpho, which creates metadata which are then stored in a file that conforms to the Ecological Metadata Language specification. The metadata librarian developed a template to use for entering metadata. Discussed if there is data that could go into a public repository.

Deliverable Three: Develop template for documenting and transferring data when students leave the lab:

- June 2016-- Created a document procedure for student archiving their data before they leave the lab at graduation. Finalized early July 2016.

Deliverable Four: Provide guidance with migrating files from older versions of Access into current versions and identify other problem files types.

- June 2016--Suggested that Project C participant hire a student intern to help her migrate old Access 97 and 2003 databases and geospatial data (ArcInfo coverages to shapefiles) into current formats (or open formats for archiving).
- March 2017—Project C participant is still intending to hire an intern to get to this, but it has not been a priority. Will follow up in three months.

Deliverable Five: Establish a method for archiving datasets that are no longer being added to, either in the campus repository or an appropriate disciplinary repository.

- March 2017—Project C participant has not been ready to do this. Concerns: sensitive data, migrate to newer format for legacy data. Discussed what kind of data could possibly be shared openly (some historical data on weather, climate, for example). However, to do this, the data would probably need to be queried and collated from old datasets that have yet to be migrated to new/open formats. As a start to the process, the participant and the project team met with the Director of the Campus Repository Service to discuss setting up a collection for the lab's projects. Project C participant suggested starting the collection with reports and white papers on the lab website, which contain some aggregated tables, and the collection was set up to receive deposits in the campus repository. The static datasets cannot be archived until she hires a student to migrate older datasets to current/open formats (see Deliverable 4). Will follow up in three months.

Project C—Exit Survey Results (see Table 3, pages 22-23) and Discussion

Project C participant felt that overall, the data management practices in her lab/group had greatly improved, 5 on a Likert scale of 1-5 where (1=not at all, 2=a little, 3=some, 4=moderately, 5=greatly) as a result of the pilot project intervention, and that the skills/information were greatly transferable to other projects in her lab. She felt that the project deliverables were met moderately overall, and she was greatly satisfied overall with her participation in the pilot. It more than met her expectations.

The three most helpful aspects of the pilot consults were the new resources she learned about; a new knowledge of metadata and the Morpho tool; and the standardized data protocols, especially for departing graduate students. She also appreciated learning about the campus repository. The pilot reinforced her view that having a data management plan (including documenting data with metadata and having a standardized data collection protocol) was valuable, and helped her to accomplish it; "Morpho was a big help with the metadata." She suggests that the consults would be most valuable when starting a project, especially with graduate students. She would also be open to attending training events offered on campus, <u>if</u> she heard about them.

Project C participant felt that data management workflow recommendations (deliverable 2) had a moderate-to-greatly positive impact on the project (and noted that she was still implementing some of the recommendations, which she expected would eventually greatly improve things). Deliverable 3, the template for documenting and transferring student data when they leave the lab, had a greatly positive impact (this was probably the most felt need). Deliverables 4 and 5 averaged a moderately-to-greatly positive impact; she was excited to learn about the campus repository and excited to start depositing reports in the collection. Once she hires an intern and gets the old data migrated to newer (and open) formats, some of that data may be deposited also, either in the Campus Repository or in the Spatial Data Explorer.

Project C participant felt that Deliverable 1, establishing iRODS as a data backup, was the biggest frustration, and had the least positive impact upon her project. She had not really seen data backup/storage as an issue; she felt comfortable with the lab server setup/backup once they moved to the new lab. She mentioned in the exit interview that she did not really see the value of iRODS currently, since they would not be using iRODS to its potential, but just as backup (in addition to their NAS box backup. The long lag time (8+ months) in getting iRODS set up and working for her may have contributed to her frustration. The utility of it may become more evident as they get all of their old data online; but even though it might exceed 5 TB, much of the old data is untouched or static. Since iRODS is not an archival solution, they are not sure what the purpose would be.

Project C Conclusions

1) Finding the points of client frustration and meeting felt needs is important. Solutions that are too general, too difficult to work with, not immediately available, require too much work, or are perceived as unnecessary, are frustrating and will probably not get implemented, or may not be used if they are implemented.

2) The more concrete and customized the help, the better. Participants wanted checklists, rubrics, protocols, and tools customized to their needs.

3) Participants implemented tools or procedures we developed for them or with them. In contrast, if we just gave guidance or told them where to look for guidance, things did not change (or were put off for a future time).

4) Participants want services at point-of-need, not before (or after)

- Example 1: since neither an expanded backup/storage solution nor a cloud-type service was a current need, iRODS and Box were not perceived as important.
- Example 2: Since they were not ready to update and migrate older data files into newer formats, the archiving goals were put off indefinitely.

Project D

Project D is a long-running (30+ years) study in cancer research with four PIs, several administrative personnel, and 5+ collaborators at three other institutions. The data consists of approximately 50 GB of questionnaire data, images, and murine models in various proprietary software packages such as SAS, Stata, Excel, Access, and SPSS. There are also physical human samples. The project data is subject to IRB and HIPAA restrictions.

Pre-pilot – status of the data

The project participants were “looking for a way to organize all historical and/or all new data in order to make it accessible for further secondary analyses” (pilot application). Data sharing had been by request only.

This project is part of a large well-funded center with several cores; the actual hardware/storage requirements were not an issue for this project.

Older datasets are organized and well-documented. The project previously had a data manager before the grant was defunded, who wrote codebooks and documented the older data before he left, including questionnaire data, images, and models in various proprietary software packages such as SAS, Stata, Excel, Access, and SPSS. They would like to make all of the old data accessible for querying and analysis.

Identified Problems with Data Management

1) Participants identified data management concerns from initial grant review, focused on Core B, to be addressed and fixed for the next internal review in the fall. This was to include an organization diagram and a research data management workflow template or diagram.

2) Participants were also concerned with study closure and data disposition details, and legal requirements for reuse or disposal of research samples. Data archiving was not so much of a concern since when the project was defunded, the former data manager did considerable work on archiving the data.

3) Since the grant had been defunded, participants were having to manage the data without a full-time dedicated data manager. The data is static and current personnel were unsure what more to do with it to archive it, and make it useful.

4) Participants wanted a data management workshop for all study personnel, including PIs, since they felt that their requests for good data management practices were being ignored.

Pilot Project D Working Recommendations

The project team will provide guidance and support to improve research data management practices for Project D participants. The project team will: create a data organization diagram and a research data management workflow; review current archiving process; identify and document grant and legal requirements for retaining data and physical samples; conduct a research data management workshop for the project PIs and staff; and make recommendations on sharing or disposal of older data or physical samples. The project participant (primary contact) will agree to meet (bi-weekly) to workshop each of the deliverables below until the developed protocols are in place and running. They will also agree to provide feedback on how the protocols are going.

Project D – Pilot Implementation timeline by deliverable

Deliverable One: Create an organizational diagram and a data management workflow for Project 3, before the September 12 deadline for a new grant proposal.

- June 2016—Met June 14 with project D participants to map out workflows and outline action points. Participants were to send several documents to pilot staff. Workflows diagrams were sketched out and finalized.
- June 30, 2016 -- postdoctoral associate leaves.

- July 2016—Project D participants sent documents to pilot staff in mid-July. Reviewed grant proposal data management plan comments via email. At the July 15 meeting discussed suggestions to improve the DMP for the grant proposal. Sent them a copy of the workflow document and a resource on how to develop a quality assurance and control plan.
- September 26, 2016 - new postdoc begins
- October 2016—unclear what the pilot participants really want in the way of a diagram or template. Review of what has been sent to them, but the time lag makes it difficult to remember what was wanted and why.

Deliverable Two: Review current archiving process, identify missing elements, and make suggestions for improvement [this was moved to a lower priority].

- October 2016 - Review of documents sent by participants regarding the previous work by the data manager. Everything is well documented and in order. It was unclear at this stage what the participants wanted to be able to do with this data.

Deliverable Three: Identify and document grant and legal requirements for retaining data, and human samples. May be supplemented by the inclusion of a librarian from UA Health Sciences (UAHS).

- November 2016—met with Marietta Marsh at IRB to discuss HIPAA requirements for de-identifying data. Met with participants and shared information; also sent them guidance from UA RDI website.

Deliverable Four: Research Data Management Workshop

- November 2016—Discussed contents of proposed workshop, created workshop content.
- December 2016—Delivered workshop and sent a copy of the slides to participants. The workshop was well-received.

Deliverable Five: Make recommendations on data sharing and/or disposing of older data and samples

- June 2016—Sent them a list of NIH-recommended data repositories.
- March 2017—gave them sample checklists for disposal or deposit of data

Project D—Exit Survey Results (see Table 3, pages 22-23) and Discussion

Project D participants were research staff responsible for managing large amounts of research data after the original grant was defunded, and after the departure of the full-time data manager. They were moderately satisfied overall with their participation in the data pilot, although they felt that the deliverables were only met somewhat overall. They felt that the pilot information, skills, and protocols were greatly transferable to other areas and projects in the lab.

They rated the lab's data management practices as "not at all improved" by the pilot. In the exit interview they clarified that the data management practices had not <u>yet</u> improved, although they hoped that practices might improve in the future as a result of the data management workshop, which was well received. The participants felt that the data management workshop bolstered their credibility with the PIs.

They felt that the workshop training the project staff provided and the referrals to websites were the most helpful aspects of the data consults. Their participation in the pilot reinforced their view of the importance of having a data management plan, and they will be more likely to complete one in the future since they are now more up to date with expectations and requirements. They suggested that, in expanding services campus-wide, that we start with RDI; get examples of successful data management plans from people on campus; and meet with large, funded grants at the beginning of the process (“initiation visits”) to get a data management plan and specific resources set up. They would be likely to attend workshops or take online/classes or tutorials if they were offered on the weekends (they are too busy during the week).

As far as project deliverables, they were most satisfied with the data management workshop the project team did for the research project PIs and staff, and with the sample checklists we gave them for dataset disposal or deposit. They were least satisfied with the data management workflow document and the review of the current archiving process. In the interview, they expressed that they were not sure exactly what they expected from the data management pilot; but that they would have liked more “hand-holding”: they wanted specific templates and checklists for their situation, rather than referral to websites for information; and specific, concrete tools rather than conceptual help. This might be, they speculated, the “gold standard” of data management support.

Conclusions from Project D

1) There were long delays in getting meetings scheduled, and in getting the documents requested from them. The more people to deal with, the harder it is to schedule meetings.
2) They wanted more support than the project team had the expertise or the resources to provide.
3) The project team did not deal with PIs, but with research staff, and this gave a different perspective on what was needed.
4) UAL needs to find/develop tools and workflows specifically to deal with UAHS clients (in conjunction with the UAHS librarians and IRB).
5) If they want the “gold standard” of support, perhaps UAL needs to adopt a fee-for-service model for the UAHS.
6) Librarians are used to providing reference services and training, but that did not seem to be enough for this client; UAL may also need to hire people with the technical expertise (or work more closely with UITS or TESS) to set up databases, or to migrate data from one set of programs to another. Projects often expect graduate students or sometimes postdocs to do this work; but the process then is often ad hoc and not well documented. For this project, all of the archived data are in (outdated) proprietary formats; to get the data in a single database to be able to query it would require exporting it all in open format, and creating a relational database, and re-importing it. This combined data is probably too large for something like Access or Excel, it requires a custom relational database in Apache/MySQL or Redcap, or tables imported into one big dataset in SPSS to query. While the Postdoc Associate had the skillset to do this work, it was outside the scope of the pilot to do this for them.

Project E

Project E is a large public health project that was fast-tracked because of its time-sensitive nature. Staff include a data manager, two PIs, four co-investigators, 40+ students and community volunteers, and a research group at another university. The data consists of environmental samples, health questionnaires, human biologic data, GIS shapefiles, and focus group transcripts, analyzed either in R or in STATA. They estimated that they had 20 GB of data, all of which was sensitive, and covered not only by HIPAA regulations but also by a tribal IRB.

Pre-pilot – status of the data

At the time of the intake interview for the pilot, the project was moving very quickly. They had not yet started the sensitive environmental or human biologic data collection, but they had already conducted a listening session and some focus groups, which were in the process of being transcribed and translated. They were planning to hire a programmer to set up a relational database in Redcap. All of the volunteers and community workers' field data were collected on paper forms, and then put into the database via dual entry with quality control procedures.

They were already using Dropbox for most of their file sharing of documentation and protocols. The GIS database was in Dropbox as well (for ease of file transfer).

Identified problems with data management

1) Document control and versioning control: with 40+ participants and graduate students, the data manager was feeling overwhelmed and having difficulty keeping track of data requests from students and the most recent copies of papers, projects, analyses, etc. in Dropbox.

2) Workflow development – wanted help creating workflow documents and data protocols before they got too far into the project.

Pilot Project working recommendations

Due to the short timeline for the project, the focus of the research data management support should be on documenting data collection, deciding on storage and back up of the different types of data, and developing a plan that can be used to prepare relevant portions of the data for either return to the tribe or destruction. The project staff should also identify potential data publishing venues.

Project E – Pilot Implementation timeline by deliverable

Deliverable 1: Document data management workflow

- May 2016 – Project deliverables agreed upon
- June 2016- Meetings about data documentation; emailed information about Open Science Framework; discussions about data naming conventions and processing and organizing focus group data; develop the environmental data workflow.
- June 30 2016 -- Postdoc leaves
- July 2016- More focus group discussion; health information workflow developed.
- Sept 26 2016 – New postdoc starts
- Oct 2016 – Data and document workflow charted; environmental and household data in Redcap. Data manager overwhelmed by document control issues in Dropbox. Analysis of document flow: recommend Trello to manage student projects, creation of final documents folder in Dropbox; restrict access. OSF considered and rejected because it does not have a stand-alone or offline version.

Deliverable 2: Establish storage and back up protocols

- Oct 2016 – Again mainly in Dropbox – need a protocol and to adopt versioning.

Deliverable 3: Develop an approach for IRB requirements, making data easier to transfer, copy and/or destroy

- Feb 2017 – Go over Tribal IRB and consent forms; legal requirements for retaining data and establishing which data and research outputs the tribe has control over; created specialized data disposition checklist for when the study is over.

Deliverable 4: Provide guidance on venues for publishing or making accessible de-identified data (such as repositories identified by NIH as acceptable places for sharing data) associated with resulting publications

- June 2016 – send list of NIH-approved data repositories for environmental data, and discuss depositing GIS data in Spatial Data Explorer. Project participants already aware of de-identification procedures.

Project E--Exit Survey Results (see Table 3, pages 22-23) and Discussion

Project E participants felt that as a result of pilot, their data management practices had improved greatly, and that the skills/protocols were greatly transferable to other projects in the lab. They felt the deliverables were met greatly as well. They were moderately satisfied with their participation in the pilot.

The three most helpful aspects of the data consults were the tools we gave them to manage the project, the workflow document, and the awareness to be more data-savvy. Their participation in the pilot confirmed for them how unwieldy and messy projects are; they learned the value of diagramming to visualize the dataflow; and it was useful to have a person research and recommend tools for them since the data manager was already overwhelmed and did not have time to do so. It also reaffirmed for them the importance of having a data management plan, but also that the plan needs to be flexible because things change. Now that they have a template to follow, they will be more likely to fill out a data management plan in the future.

Their biggest frustration with the pilot was the speed at which the project moved (it was fast-tracked); our advice was often out of date before it could be implemented. The 3-month delay between when the first postdoc left and the second was hired (end of June-end of September) missed a crucial part of the process; they needed us to work more closely with them at the beginning.

They had several suggestions for scaling up services to campus. They suggested that since the UAHS are used to a fee-for-service model, perhaps the library could scale up services by charging a fee for service. The PI said she would be unlikely to attend any data management classes, since she has no time. The data manager said it would be difficult for him as well; but they said it would be beneficial for graduate students. They said the library used to offer a one-unit literature review course, and the students loved it: they urged us to bring it back! They also suggested that the library offer a for-credit data management class. They suggested training for new graduate students, for new faculty, and at the beginning of new grants.

In their assessment of the project deliverables, they felt that documenting the data management workflow had a greatly positive impact on their data practices. The storage and backup protocols had a moderately-to-greatly positive impact. Developing an approach to the IRB requirements and data disposition had a moderately positive impact. Deliverable Four, providing guidance on depositing de-identified data in a suitable repository, was not applicable because the tribal IRB dictated that all data be destroyed or be sent back to the tribe.

Conclusions from Project E

1) On fast-tracked projects, data services need to be agile and continually available at point of need, and clients would be willing to pay a fee for service model to get that kind of support.

2) PIs do not feel they have time to attend data management training, and staff feel too busy as well. PIs and staff feel that graduate students need such training, but suggest offering it as a for-credit class for them to make it worth their while.

3) Offer training and support services at new faculty orientation and at the beginning of grants.

4) Faculty and staff need some basic instruction in how to use cloud services like Dropbox or UA Box for document control, versioning, and file sharing. Whatever solution they have must be able to work offline as well, so that they can work in the field without an internet connection; this requirement might preclude using Open Science Framework, since OSF has decided not to develop a desktop version. However, field researchers could use cloud services like Dropbox and attach those files to OSF to mitigate this limitation.

Table 3: Positive Impact of Deliverables by Project (note that in this table, deliverables are not comparable across projects. These results are only for projects for which we had exit surveys.)

Deliverable	**Project C – Wildlife Biology**	**Project D – Cancer Research**	**Project E – Public Health**
1	Establish iRODS data storage as backup (ranked 2 on a Likert scale of 1-5)	Document data management workflow (ranked 2 on a Likert scale of 1-5)	Document data management workflow (ranked 5 on a Likert scale of 1-5)
2	Data management workflow (ranked 4 on a Likert scale of 1-5)	Review current archiving process (ranked 2 on a Likert scale of 1-5)	Establish storage/backup protocols (ranked 4 on a Likert scale of 1-5)
3	Template for documenting student data when they leave the lab (ranked 5 on a Likert scale of 1-5)	Identify and document grant/legal reqs for retaining physical samples (3 on a Likert scale of 1-5)	Develop an approach for IRB requirements to make data easier to transfer, copy, or destroy (4 on a Likert scale of 1-5)
4	Guidance for migrating files from older versions of Access (ranked 4 on a Likert scale of 1-5)	Data management workshop for all PIs and research staff (ranked 5 on a Likert scale of 1-5)	Guidance for making data available to publish (N/A because of Tribal IRB)
5	Method for archiving finished datasets (ranked 5 out of	Make recommendations on sharing/disposal of older datasets (ranked 4	N/A

	on a Likert scale of 1-5)	out of 5 on the Likert scale)	
"Avg" Likert ranking	4.0	3.2	4.3

Pilot Project Statistics

The following statistics are for person-hours spent working on the implementation phase of the pilot, and not on the planning or assessment phases. As such, they are estimates of the hours actually spent consulting with research projects and working on deliverables. The project team kept track of face-to-face consult hours and staff meeting hours via calendar appointments. Individual hours of working on projects are [low] estimates based upon project team notes and emails.

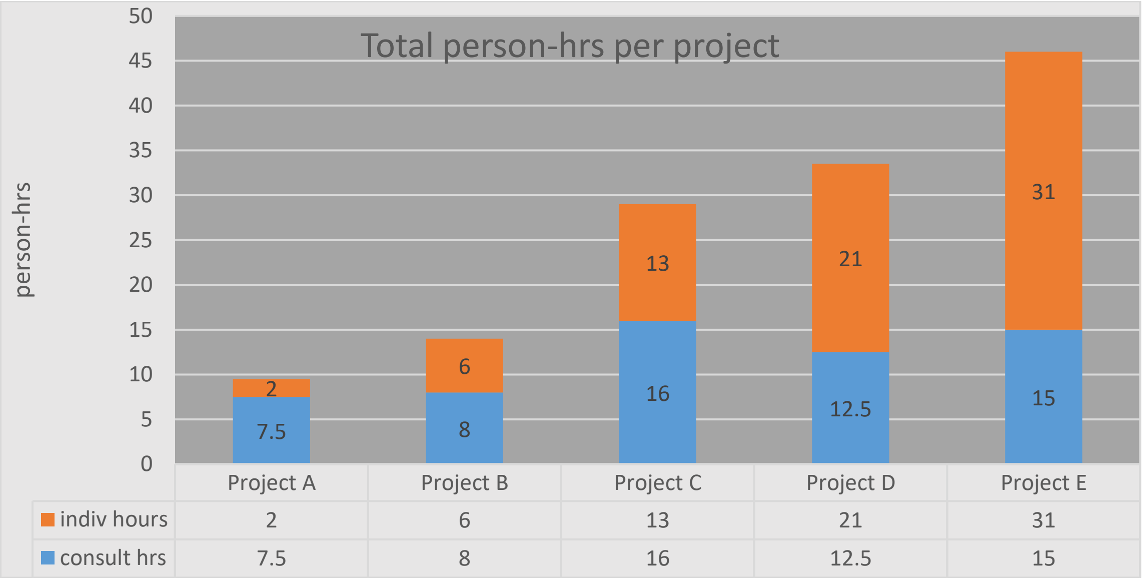
Overall time spent

The project team estimate that 35% of the total hours spent on the pilot were in face-to-face consults with the researchers; 21% of the hours in meetings with each other to discuss the projects and strategize deliverables; and an estimated 44% of the hours working independently on tasks related to the deliverables, for a total of 166.5 hours.

Time broken down by project

The total person-hours per project ranged from ~10 hours for the smallest project to ~45 hours for the largest, most complex project, plus an average of 7 hours of staff meetings/project. See Figure 3 for comparisons.

Figure 3 Time Broken Down by Project



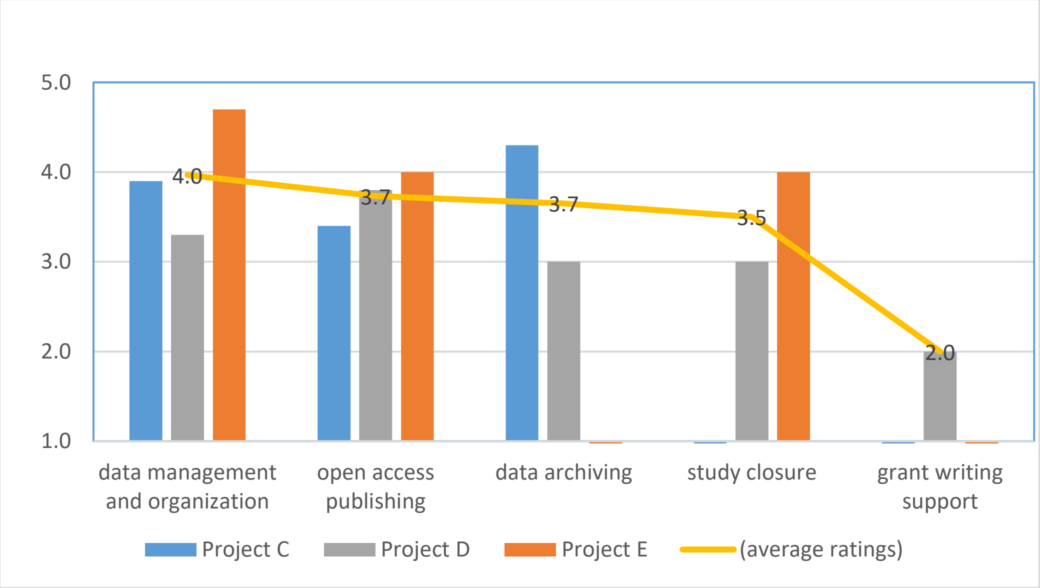
Improvement ratings

For the three projects for which exit interviews were conducted, participants were asked about the degree to which participation in the pilot improved their data management practices, on a Likert scale from 1-5 where (1=not at all, 2=a little, 3=some, 4=moderately, 5=greatly). Broken down by topic, the largest improvement rating was in data management and organization, ranging from some to greatly, with an average improvement rating of moderate. Open access publishing, data archiving, and study closure average improvement ratings were in the range of some to moderate improvement.

Table 4: Summary of Exit Interview Improvement by Topic

Topic Category	Project C	Project D	Project E	Average ratings
data management and organization	3.9	3.3	4.7	4.0
open access publishing	3.4	3.8	4.0	3.7
data archiving	4.3	3.0	n/a	3.7
technical skills	n/a	n/a	n/a	n/a
study closure	n/a	3.0	4.0	3.5
grant writing support	n/a	2.0	n/a	2.0

Figure 4 Data Management Improvement by Topic



The only study variables that correlated highly with the total pilot staff hours spent on each research project were the number of people involved in the research project, including PIs, collaborators, graduate students and postdocs. Legacy data status, protected data status, subject area, amount of data, the length of the project, or the number of deliverables did not correlate with the total hours that pilot staff worked on a project.

Discussions and Recommendations

For the data pilot, more than half of the project deliverables can be categorized as data management (51%). Another quarter of the deliverables were in support of open access services (26%). Data archiving was a distant third at about 10%. Data management included basic help with the best practices of data workflow management: data gathering, data transfer, data documentation, workflow documentation, file naming and organization, data and document sharing within the research group (mainly between students and PIs), working file storage, file backup, and to a lesser extent, writing and/or updating a data management plan. Open access services included providing information about depositing data supporting publications into the appropriate repositories. If the pilot projects are representative of campus projects as a whole, then the university needs to expend significant resources in training and supporting faculty, research staff, and graduate students/postdocs in their daily data management tasks and workflows.

Data archiving included preserving datasets long-term, and making them available in open data formats, either in the UA Repository or in an appropriate disciplinary data repository. However, none of the research projects were at a point where this step was actually tenable for their data: they needed more support in preparing data for archiving than the pilot personnel could provide. For example, they required assistance in de-identifying sensitive data, updating legacy data into a current proprietary format (like Access) so that it could be worked with; in selecting data to archive; in exporting data into open formats. One had tribal IRB restrictions that mandated returning all data to the tribe and destroying any copies. Moreover, our current campus repository is not currently able to accept large data sets or sensitive data.

General recommendations:

- Support needs to be a joint effort between the Libraries, UITS, and RDI.
- Data management and curation services need to be fully developed, robust, agile, and available at point of need.
- Researchers will not necessarily reach out, so data management support needs to be **offered** at time of need, preferably in a fashion that is already part of the researcher's workflow.
- Researchers feel overwhelmed and want to devote as little mental bandwidth or executive function as possible to the tasks and tools surrounding data management. Ideally, it should be as transparent as possible to the researcher, and integrated into existing workflows.
- Researchers desire customized support that works with tools the researcher is already familiar with (as much as possible).
- These services should be unified as much as possible, at least from the researcher perspective, so that they do not have to contact ten different offices or departments to get services. Investigate the possibility of developing a unified dashboard or portal.
- Concrete solutions are better: researchers wanted checklists, rubrics, templates, protocols, and tools customized for their needs. Just sending them links to information or resources did not insure that they understood or implemented the information.
- Data management best practices are especially crucial the larger the research group; intra-team communication practices, shared data storage and protocols, and version control practices are especially important. It is also crucial to have continuity of data management practices – especially for research projects that hire post-docs and graduate students (data documentation is critical).
- Most research on campus is not "big data" research, and may not need HPC resources. However, they need secure working storage and a long-term archival solution.
- Offering workshops on data management best practices are a good first step, but those practices are difficult to transfer to the lab without PI buy-in, and without a consultant to help

document and structure the data workflows, since it is unlikely that PIs will attend the workshops themselves.

- For UAHS researchers, HIPAA requirements make data management and disposition more complex; they may be more willing to accept a fee-for-service model than main campus researchers.

Specific recommendations:

- A suite of services should be offered at point of need during different points in the research lifecycle.
    - At the beginning of research projects: provide assistance with grant proposals and the data management plan, the allocation and organization of working data storage, data collection and management protocols, file naming conventions, setting up a lab wiki or google/dropbox/box account for projects with sensitive data). Perhaps also setting up research databases or Redcap.
    - During active data collection and processing: support for evolving data needs, updating the data management plan, assistance with data documentation and version control protocols. Provide assistance with learning how to do data cleaning, analysis, and use of visualization tools.
    - At the publication phase, assistance in de-identifying and reformatting data for deposit with a manuscript.
    - At study close out, assistance with data disposition according to IRB requirements.
    - At the archiving phase, assistance in metadata and conversion of data into open formats for long-term storage, and assistance in deposit in the campus data repository or a disciplinary data repository.
- Offer a service like Open Science Framework (OSF) for Institutions, and encourage use by offering 20 hours of data management consultations to grant startups that agree to use OSF as their shared data storage and communication system.
- Coordinate services with the Office of Sponsored Projects and other units so that the library is notified when grant proposals are submitted/and or funded, or at various project review points, which could automatically trigger an email to researchers offering data management services and resources.
- Investigate campus partnerships with UITS, RDI, and CyVerse in developing a data repository.
- Develop a campus data repository – provide management, sharing, and preservation of research data in a structured infrastructure. As noted earlier by Erway and Rinehart (2016), research funders are looking to institutions to provide research data preservation solutions rather than access nodes being provided by the majority of disciplinary data repositories. Overall advantages for the University include:
    - Achieve efficiencies by having a central repository on campus for researchers to deposit their data.
    - Data will be more accessible and findable
    - The University will see a competitive advantage, funding agencies will be more confident in the resources and infrastructure of the University, potentially resulting in a higher percent of successful grant proposals.

  Current options for UA researchers include the Campus Repository, disciplinary repositories, and the Spatial Data Explorer (geospatial data). The Campus Repository and disciplinary data repositories have different advantages but also several challenges and issues:
    - The Campus Repository can handle all formats, although it is not designed for data or an ideal solution. In addition, there are limitations on file size (<300 MB), it cannot handle sensitive data and it is difficult to provide different levels of access to files.

- Disciplinary data repositories' advantages include disciplinary data is together in one repository, may be able to handle sensitive data and may include discipline specific analysis tools. Disadvantages include limitations on file format and file size, may not provide preservation services, and the long-term sustainability of some data repositories is questionable.

- Investigate electronic lab notebooks to use as a collaboration platform. Conduct survey of what is currently being used, what needs do researchers have. Is there an ELN that the campus should support?
- Add staffing, training and funding for liaisons librarians, who could suggest discipline-specific resources to help with data discovery and archiving, and who conduct departmental data management workshops and consultations.
- Continue and expand training opportunities for graduate students and postdocs to acquire skills in data management and curation in their discipline. Consider developing an online for-credit course and working with instructors (in conjunction with liaisons) to integrate data management into discipline-specific courses at the graduate level.
- Offer Data Management training to new faculty and graduate students as part of the Responsible Conduct of Research training.
- Develop online data management tutorial modules.
- Update and expand the UA Data Management Resources website
- As the Data Curation Network (DCN) is implemented, consider how their data curation services could supplement and enhance the services we offer at the UA.

Table 5 – Pilot Recommendations with Priority and Responsibility

Service or Project	**Priority Level**	**Responsibility**
Open Science Framework for Institutions	A	Libraries (ODIS) and UITS
Coordinate services with RDI, Sponsored Project; notifications of library's RDM services	A	Libraries (ODIS), RDI (Sponsored Projects)
Data Repository	A	Libraries (ODIS and TeSS), UITS?
Training for liaisons (R & L, UAHS)	A	Libraries (ODIS)
ELNs	B	Libraries (ODIS)
Training sessions for graduate students and post-docs	B	Libraries (ODIS and R & L)
Online Tutorials	B	Libraries (ODIS, R & L?), RDI?
Data Management Website update	C	Libraries (ODIS and TeSS)

References

Erway, Ricky and Amanda Rinehart. I*f You Build It, Will They Fund? Making Research Data Management Sustainable*. Dublin, Ohio: OCLC Research, 2016. http://www.oclc.org/content/dam/research/publications/2016/oclcresearch-making-research-data-management-sustainable-2016.pdf

Fearon, David Jr., Betsy Gunia, Barbara E. Pralle, Sherry Lake, and Andrew L. Sallans. *SPEC Kit 334: Research Data Management Services*. Washington, D.C.: Association of Research Libraries, 2013. http://publications.arl.org/Research-Data-Management-Services-SPEC-Kit-334/

Holdren, J.P. *Increasing Access to the Results of Federally Funded Scientific Research*. Office of Science and Technology Policy, Executive Office of the President. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Hudson-Vitale, Cynthia, Heidi Imker, Lisa R. Johnston, Jake Carlson, Wendy Kozlowski, Robert Olendorf, and Claire Stewart. *SPEC Kit 354: Data Curation*. Washington, D.C.: Association of Research Libraries, 2017. http://publications.arl.org/Data-Curation-SPEC-Kit-354/

Johnston, Lisa R., Jake Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert Olendorf, and Claire Stewart. *Data Curation Network: a Cross-Institutional Staffing Model for Curating Research Data*. 2017. https://conservancy.umn.edu/handle/11299/188654.

Provost's Task Force on the Stewardship of Digital Research Data. *Research Data Stewardship at UNC: Recommendations for Scholarly Practice and Leadership.* Chapel Hill, NC: University of North Carolina, 2012. https://sils.unc.edu/sites/default/files/general/research/UNC_Research_Data_Stewardship_Report.pdf.

Wilkinson, M.D., et. al. *"*The FAIR Guiding Principles for Scientific Data Management and Stewardship". *Scientific Data* 3:160018. https://www.nature.com/articles/sdata201618.

APPENDIX

Data Management and Data Publication and Curation Pilot – Final 4/21/2015
Developed by the RCGC Data Management and Curation Subcommittee

Introduction

Research carried out at the University of Arizona should have the widest possible impact. Archiving data outputs from federally funded research has been mandated by several agencies and the majority of new federal grant proposals require a cogent and achievable data management plan. To ensure impact, data outputs from University research should be available in a robust, discoverable archive. As federal funds for research have declined, technology to collect, store and analyze data has improved significantly. Data and workflow management have become a challenge for research groups worldwide, without tools to effectively manage data, the benefits of improved data collection and analytical techniques may not be fully realized.

The development of an efficient set of data management and curation services that include training and support for researchers and the needed technology infrastructure for these services will allow researchers to devote more time to research, support the goal of developing a research data ecosystem that facilitates data reproducibility, and have a positive impact on the UA's overall prestige and success in obtaining grant funds.

Purpose and Goals

The following proposal presents the structure and funding needs for a Data Management and Data Publication/Curation pilot. The pilot will work with 3-5 research projects through the entire data lifecycle to evaluate the implementation of data management services at the UA. The pilot will develop what services will be needed, evaluate what worked well as well as what didn't work well, training needs, feasibility of implementing data management services for campus, and demonstrate success.

Milestones

Milestone 1: Identify 3-5 suitable research projects. Work with the Office of Research and Development to identify research projects that have recently (within the past 6 months to 1 year) been awarded grant funding. The research projects should span different data types, scale of research data to be collected or created, and size of the research team (one researcher to a large research group). Evaluate what level of support each project will need.

Milestone 2: Update and implement each research project's data management plan; set up data management technology; recommend the data workflow and appropriate metadata standard to use for describing each data type. Check in with researchers on a regular basis.

Milestone 3: After 5 months, evaluate utility of data management tools and services recommended for each project. Work with each research project to improve utility of tools and services or recommend new tools.

Milestone 3b: Survey who is using Box and Google Drive for research, determine how widespread the adoption of these technologies is on campus.

Milestone 4: Work with researchers on getting the research data 'repository ready' -- descriptions for shared data and tools (such as software).

Milestone 5: Evaluate the pilot and provide recommendations – what are qualitative observations from researchers and service providers? Was the pilot a success? What were the cost savings? What were the training needs, how much face-to-face time was needed? What active data management tools are recommended? What's the feasibility of ramping these services up to the entire campus and what would be needed – technical infrastructure, tools, staffing needs etc.

Services - Detail

1) Data management service for active data storage could offer a toolbox of technology products that faculty could choose depending on what would easily integrate into their research workflow.
    a. Possible products include the following:
        i. Box.com – we currently have a UA implemented Box.com occurrence. It provides up to 5GB for individual files, 35 GB of total space; provides the ability to share with others working on a research project, has file versioning, and the ability to add comments to files. In addition, UITS is currently looking at implementing a HIPAA compliment version of Box.com for the UA.
        ii. Google Drive – we currently have a UA implemented Google Drive as part of UA Google Apps for Education. It provides up to 30 GB of total space, the ability to share with others working on a research project, has file versioning, and the ability to add comments to files.
        iii. iRODS -- is policy-based data management middleware software that maps between protocol from the client to protocol from data storage and was developed by University of North Carolina's Renaissance Computing Institute (RENCI). iPlant has deployed iRODS and are willing to work with UITS to install an iRODS server on UA computing systems. They also have a close relationship with RENCI, if we have any issues or questions.
2) Data publication and curation service will collect, share, and archive research data created by UA researchers and their collaborators, providing persistent long-term access. Researchers will be able to publish their research data with persistent identifiers that other scholars can use to find and cite their data. The service could potentially use Rosetta. Rosetta was recently purchased by the UA Libraries as the University's preservation repository system and includes a discovery interface. The Libraries will be taking the lead in implementing Rosetta for campus including developing policies and workflows. This assumes that any data or software that a researcher develops as part of a research project and wants to share, archive or preserve will be made available either in a disciplinary data repository or in a UA data repository and not both.

Cost Estimate

Servers and storage estimate-- $70,000
The configuration for a pilot instance of iRODS would involve 3 servers (total of $45,000) and 60TB of usable storage added to our existing research storage array ($25,000). We are planning to incorporate this iRODS instance in the science DMZ to support high speed data transfers. We can start with 2 servers and add the third when the usage exceeds capacity. (Note: If we want stand-alone storage, not part of the existing HPC environment that will need to be estimated differently.)

Staff servers and storage support estimate -- $7,400 (the cost to support a server is $1850, and counting the storage as another server).

Data management and curation staffing support estimate -- $100,000 (embedded data management librarian in 3-5 research projects, developing templates, workflows etc.; programming support; metadata support)

Total cost -- $177,400

Background

The RCGC Data Management and Curation Subcommittee administered a research data management survey to faculty, researchers, and graduate students during spring 2014. One of the questions listed 14 current or potential data management and curation services and asked participants to indicate if they were interested in the service. The services with the highest interest were data storage/management and a data repository. These are two services that clearly need to be addressed at the campus level.

Supporting documents – RCGC and UITS

The pilot supports the goals of the RCGC Data Management and Curation Subcommittee and the RCGC Big Data Subcommittee to "…identify campus wide big data requirements and **formulate priorities** for the institutional support of big data, develop policy around institutional support, and **make recommendations** on funding institutional support of big data." In addition, the University of Arizona Campus Cyberinfrastructure Plan under the heading: Improve Institutional Efficiency lists as an upcoming activity, evaluation of a campus wide data management infrastructure using iRODS.