Los Alamos
NATIONAL LABORATORY
EST. 1943

LA-UR-17-21959
Approved for public release; distribution is unlimited.

Title: Data Management Services at Los Alamos

Author(s): Finnell, Joshua Eugene
Cain, Brian J.

Intended for: Report

Issued: 2017-03-07

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

RESEARCH DATA MANAGEMENT SERVICES AT LOS ALAMOS NATIONAL LABORATORY

A Research Library Report on Data Management Needs & Services

Prepared by the Research Library *Data Working Group*:

Joshua Finnell, Brian Cain,

Kelly Durkin, Carol Hoover, Dee Magnoni

September 1, 2016

Table of Contents

Executive Summary

This data management report was commissioned by the Research Library Data Working Group for two purposes:

1.) Conduct an environmental scan of external institutions to benchmark budgets, infrastructure, and personnel dedicated to data management.
2.) Perform a survey of the data management landscape at Los Alamos National Laboratory in order to identify local gaps in data management services.

The first stage of data collection consisted of contacting data librarians and managers at 12 institutions, universities and national laboratories, during the spring of 2016.

The second phase of data collection consisted of 24 in-depth interviews with researchers from across the Lab and were completed during the summer of 2016. The individuals who were interviewed spanned a diverse set of divisions, positions, and career stages but should not be considered comprehensive. It should also be noted that previous data collection in the area of data management has been undertaken at the Lab in the last five years. Surveys were conducted by the Research Library in 2011 and by a previous iteration of the Data Working Group in 2015. Also in 2015, Reid Priedhorsky completed data management interviews with members from High Performance Computing. These surveys and interviews provided both historical context and guideposts in conducting this current version of data interviews. Moreover, data management needs remain consistent across the last five years of surveying: an understanding of data management plans, centralized collaborative tools, and data storage.

Key findings focus on identified data management services at Los Alamos National Laboratory as well as benchmarks for data infrastructure, cooperation, and staffing.

Institutional Benchmarking

– The majority of institutions surveyed have an average operating budget of $500,000 annually dedicated to data management services. These budgets support on average 3 full-time employees and a data platform such as Dataverse or DSpace. Collaboration was a common theme across institutions, with data services being executed collaboratively between 3 or more departments.

Data Management Planning

– The awareness and completion of data management requirements and mandates was limited in our survey pool, even with the Department of Energy's Office of Science mandate requiring data management plans (DMPs) for federally-funded research on October 1, 2015. The few interviewees who created a data management plan used the library-sponsored DMPTool, but better support and services were requested by researchers in this area.

Data Storage – Echoed in the data surveys conducted in 2011, the lack of a centralized data storage solution at the Lab, that meet the needs of the research community, was a common theme. For myriad reasons, from mere efficiency to cost, the most commonly employed approach to data storage and preservation is a personal computer or local network drive.

Data Collaboration and Dissemination – Connected to the issue of centralized storage is a desire among many respondents for collaborative tools (i.e. Google Drive, Dropbox) to work with lab partners on research projects. Currently, LANLTransfer or email attachment is the preferred method of sharing data with researchers both internally and externally. Additionally, the majority of researchers requested assistance with submitting their data through RASSTI as the current system requires copying data to a physical CD and delivering it to SAFE-1 for review.

Though time and the competitiveness inherent in publishing and securing grants was often cited as a challenge to data management by researchers at the lab, the continued absence of a centralized data repository and effective collaborative infrastructure is an obstacle in fulfilling funder and publisher mandates and conducting scientific research efficiently and effectively. The Research Library, providing research support for every division and department across the Lab, is uniquely capable of building infrastructure to manage the data lifecycle, from creation to dissemination to preservation. However, these services require appropriately trained support personnel, collaboration with allied departments, and sustainable funding models to meet the expectations of researchers and federal mandates. A full data infrastructure proposal can found in Appendix 1.

Introduction

In the last decade, scientific research has become more data-intensive and collaborative. As a corollary, researchers at Los Alamos National Laboratory need more assistance organizing, storing, publishing, archiving, and sharing their data. Almost every major funding agency, from NASA to the NSF, has established or will require a robust data management plan to secure and maintain research funding. In addition, scientific journal, such as Nature, are also requiring researchers to make the data underlying their research paper publicly accessible.

Currently, the Research Library at Los Alamos National Laboratory offers data management services within the planning phase of the data management. In addition to assisting researchers

in understanding federally-funded data mandates and publisher requirements, the library also provides ready-made data management plan templates for major funding sources. Maintaining currency with both discipline-specific and general data depositories, the library staff is knowledgeable in identifying and discovering relevant data sets for lab employees to integrate into their research. However, planning and discovering are just the beginning of the research data management lifecycle. Storage, sharing, publishing, and archiving one's own research are also crucial stages in making research data findable, accessible, interoperable, and re-usable (FAIR) to the scientific community. The purpose of this report is to assess the needs of researchers across the Lab in data management, and to inform the development and coordination of services needed in the Library to support effective data stewardship through the research lifecycle.

Institutional Data Benchmarking Survey

Before exploring the data management landscape internally, the Data Working Group was tasked with providing an environmental scan of data management services at universities and national laboratories identified as *leaders* in developing data services and infrastructure. As a first step, the Data Working Group identified two reports in the professional literature: *The International Survey of Academic & Research Library Data Curation Practices, 2016-17 Edition* & *The Association of Research Library's SPEC Kit 334: Research Data Management Services*. From these surveys, and in conjunction with the Data Executive Committee, the Data Working Group identified 12 institutions and contacted data services managers at each institution. Data gathered from Purdue University was culled from their two-day visit to the Research Library in December 2015 to specifically discuss their approach to data management services and infrastructure.

Interviews with each institution were conducted over the phone with the following questions:

- Budget {high level} - what is included in this figure/excluded?
- Staffing - current total # FTE, # fulltime/part-time if available, # professionals versus others if available (i.e. librarians, IT staff, administrative, etc.)
- Data Security/Platform - how is data security ensured - what mitigations or software tools are used? What is the repository software platform? Name of repository?
- Discovery Tools - what tools are used to find/locate data content in the platform?
- Interdepartmental Cooperation - internally or externally, who are they collaborating with to provide research data services and/or the repository platform? Note this does not include the researchers who deposit/use the data.
- Permanent Data Storage - how are they handling data preservation (permanent storage of data)? Or are they not providing permanent storage? Are there limitations on storage time? Fees?

The responses were compiled and summarized into common themes and averages to provide a benchmark of expectations for administrative and budget support for similar data services at the Research Library.

The full survey and summary table are available in Appendix 2.

Budget	$200,00 - $2,000,000
Staffing	2-6 FTE
Platform	Hydra/Fedora + Dataverse + DSpace
Data Discovery Tools	DOIs + APIs
Interdepartmental Cooperation	Partnership between 3+
Permanent Storage	10-year retention policy

Data Interviews

In addition to an external environmental scan, the Data Working Group also conducted an internal environmental scan of researcher's data management needs. To select participants, the Data Working Group employed five strategies:

1.) Identify all LANL data sets deposited in external data repositories by LANL researchers (Figshare, Zonodo, Dryad, Dataverse)
2.) Contact all researchers who have submitted a data set through RASSTI
3.) Contact all researchers who have created a Data Management Plan using the Library's DMPTool
4.) Identify data-intensive researchers from the Library Roadshows
5.) Recommendations from the Data Executive Team.

In total, approximately 102 potential interviewees were identified during this environmental scan, and interview subjects were primarily drawn from this pool. Each researcher was emailed individually and invited to meet with a member of the Data Working Group for a brief 30-minute data interview. It should be noted that the Data Working Group intentionally targeted "smaller" data sets at the Lab, as big data is currently beyond the scope of the Research Library.

Participants

The Data Working Group aimed for a diverse pool of interviewees, both in terms of research rank and division. However, due to response rate, the data in this report should not be considered a random sample. Findings should not be interpreted as quantitatively representative of any demographic group identified in this study. In total, 24 interviews were conducted.

Design

Individual interviews were kept to 30-minutes and discussion was semi-structured, based on a prepared list of questions around four major areas of data management: planning, storage, publishing, and policy. The interview questions were developed by the Data Working Group and initially tested and validated with a focus group of researchers in EES-16: Computational Earth Science. In addition, the Data Working Group worked with Ben Simms in CCS-6: Statistical Sciences to further refine the interview questions for precision and accuracy. Ultimately, the interview questions employed a mixed methods approach, integrating both quantitative and qualitative questions.

Though some questions were required, interviewees were not expected to answer every question. The goal for qualitative inquiry was to encourage researchers to elaborate on areas of concerns or "pain points" in the data management lifecycle. There responses are reflected in the data interview findings.

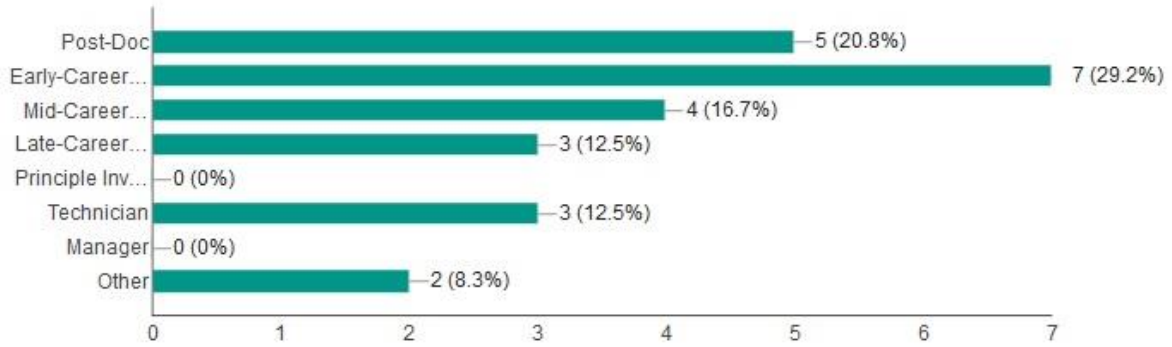
As previously mentioned, the purpose of this study was to identify potential service areas for the Research Library in the data management lifecycle at the Lab.

To protect the identity of participants, personally identifiable information has been removed from the results.

A full list of interview questions can be found in Appendix 3.

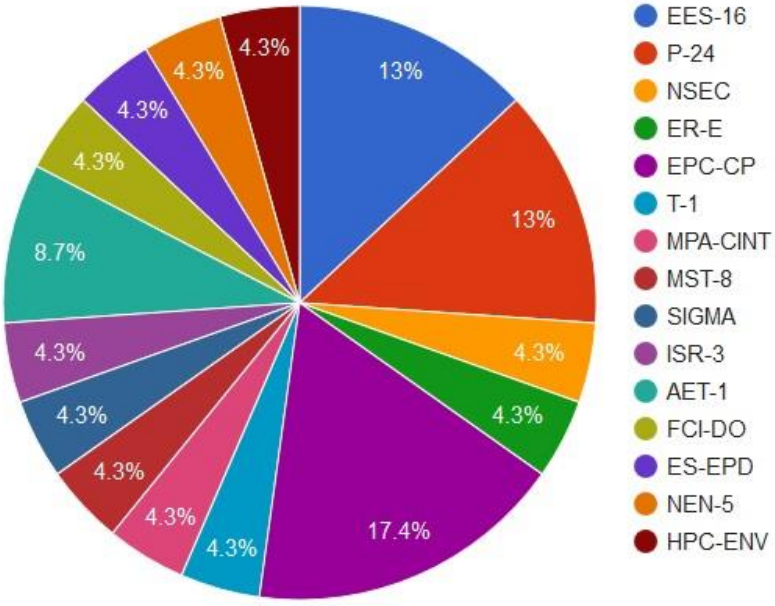
Data Interview Findings

Demographics



Employee type roughly demarcated using the Lab's designation by age and type of research conducted. For example, early-career researchers held non-postdoctoral research appointments and were under the age of 35. Technicians were identified as those working in the fields of compliance, as opposed to traditional research.

Of the 24 participants, almost half were post-docs or early-career researchers. The response from this demographic is reflective of recent data mandates affecting mid-career researchers and the increased education on, and importance of, data management in doctoral programs.



Though evenly distributed, EES-16: Computational Earth Science and EPC-CP: Environmental Compliance are more heavily represented for two reasons:

1.) Initial validation of the data interview questions was conducted with three members of the EES-16 team.
2.) Due to mandates requiring public release, technicians and researchers in EPC-CP submit all of their data through RASSTI.

Overall, the diversity of divisions represented underscores the breadth of data management needs across the Lab:

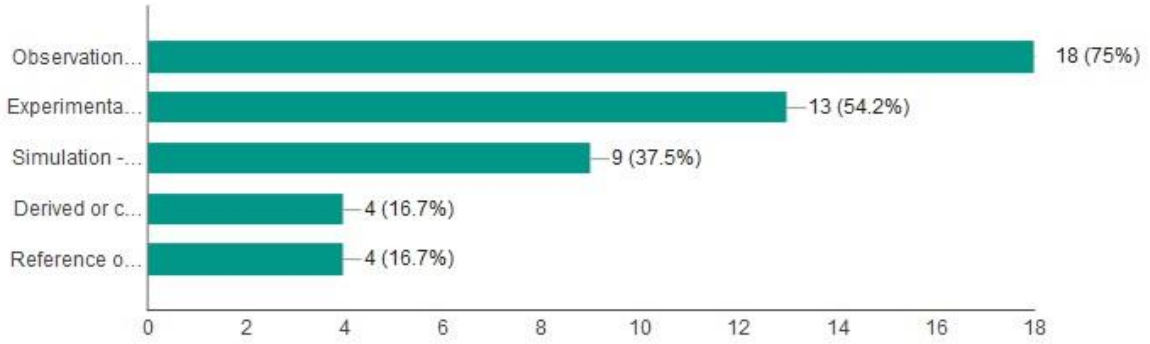
> "Working on a project to change water quality standards for acceptable temperatures for discharges into high quality cold water areas with aquatic life. There is a temperature requirement for discharges into the waters of the US that we have to meet and we can't meet it. Using sensors deployed in canyons where water is flowing to measure dissolved oxygen content and temperatures. Data is collected every 15 minutes."

> "Study phase transformations in titanium and other structural metals using x-rays. Work with APS (Argonne) and will more than likely have a role with MaRie."

"Flow and transport simulation modeling – flow of fractures in the rock (need fracture baseline characteristic) – data is hard to get. We need historical data at nuclear waste site in shale (data is proprietary) and fracture characteristics for any rock types."

"We work with Hyper V Technologies and a multi-institutional team to develop a plasma-liner driver formed by merging supersonic plasma jets produced by an array of coaxial plasma guns."

Data Types & Formats

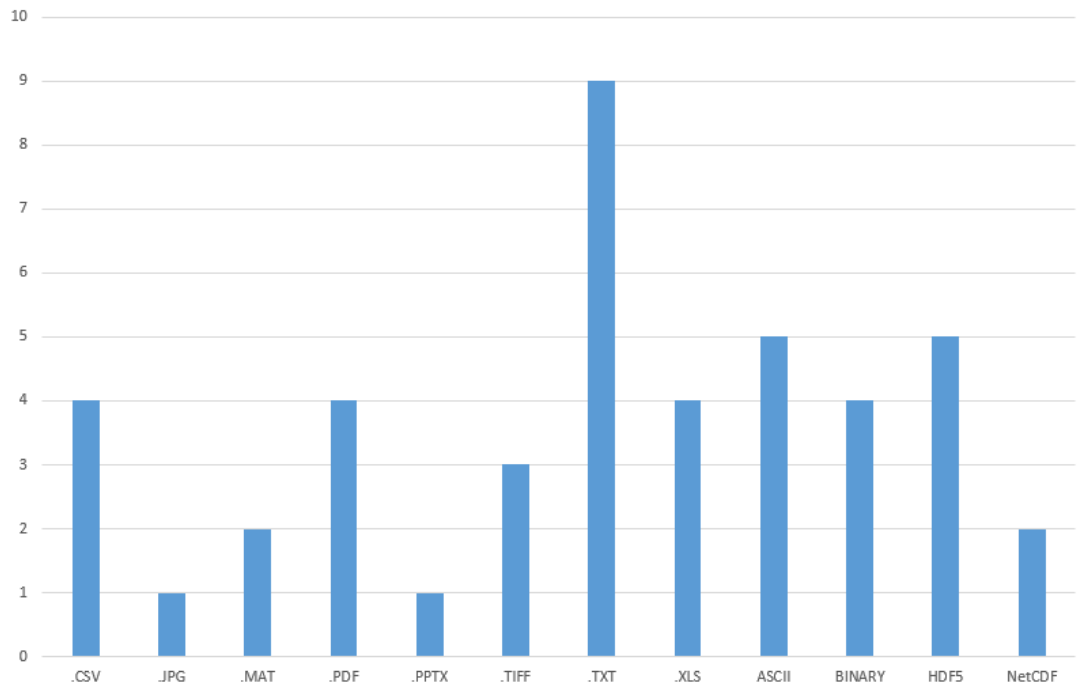


Derived from the institutional benchmarking survey, and professional organizations such as *Data One*, data types were categorized as follows:

- **Observational**: data captured in real-time, usually irreplaceable e.g. sensor data, survey data, sample data
- **Experimental**: data from laboratory equipment, often reproducible, but can be expensive to reproduce e.g. gene sequences, chromatograms, toroid magnetic field data
- **Simulation**: data generated from test models where the model and metadata are more important than output data e.g. climate models, economic models
- **Derived or compiled**: data is reproducible, but expensive e.g. text and data mining, compiled database, 3D models
- **Reference or canonical**: a (static or organic) conglomeration or collection of smaller (peer-reviewed) datasets most probably published and curated e.g. gene sequence databanks, chemical structures, spatial data portals

The large number of data types is representative of the diverse and blended research conducted by the interviewees, with many participants working with multiple data types. The predominance of observational and experimental data is attributed to:

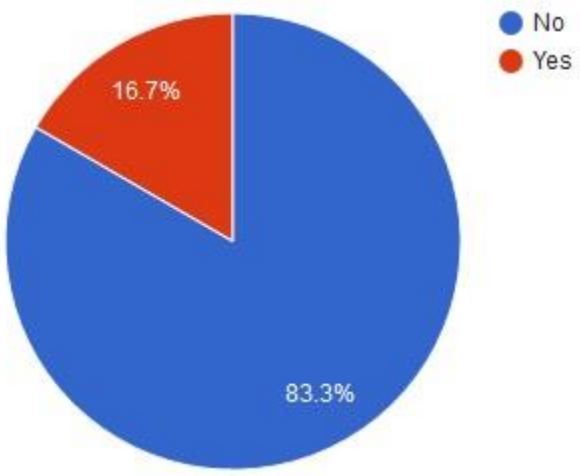
1.) The number of participants interviewed from the environmental compliance division where monitoring and sensors are heavily employed.
2.) Reflection of the Data Working Groups focus on “smaller science.” The majority of interviewees identified their data sets as less than 10 TB



The “small science” is reflected in the data formats mostly commonly used by the interviewees. The use of non-proprietary formats among interviewees is a promising sign of reproducibility. However, the small presence of .pptx and .xls files suggest an educational outreach opportunity for the Research Library. The prevalence of HDF5 and NetCDF, both portable and extensive file formats and models that support an unlimited variety of datatypes, among the interviewees uncovers a potential file type around which a data infrastructure could be designed.

> “Our goal is to roll all file formats into 1 format (HDF5).”
>
> “We compress everything (.pdf, .txt, .tiffs) into HDF5.”
>
> “All of our data is about 1-2GB total.”
>
> “We work with small data. I’d say less than 1 GB.”

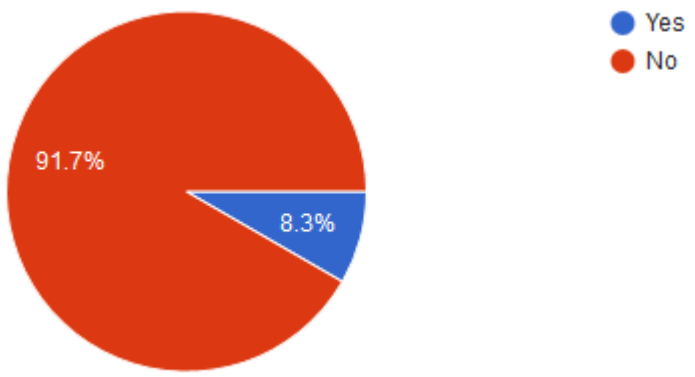
Data Management Planning



Data Management Plans (DMPs) are supplementary documents that outline how a researcher will ensure data is findable, accessible, interoperable, and reproducible during and after the research process. DMPs are currently required by almost all major funding agencies. The overwhelming majority of interviewees have never created a data management plan. The small percentage of researchers who have created a data management plan were those whom the Data Working Group identified through the DMPTool, a service provided by the Research Library. All 4 respondents who created data management plans were required to by their funding agency. The responses reflect the newness of funder and publisher mandates, and also demonstrate an educational outreach opportunity for the Research Library with the DMPTool and LDRD proposals.

> “Well, the grant was funded before any mandates. We don't currently have a data management plan.”
>
> “Our current grant required a DMP from the Office of Science. So we wrote up a DMP from their website.”
>
> “Phrase ‘data management plan’ isn’t meaningful. Required to submit a "data work plan" to the state that identifies data gaps, how we will collect the data, and other information that will eventually lead to a change in water quality standards.”
>
> “No data management plan, only follow lab policy.”
>
> “Honestly, the link to the DMPTool through LDRD wasn't clear. Wasn't sure how to fill out the DMP to satisfy LDRD Requirements.”

Metadata

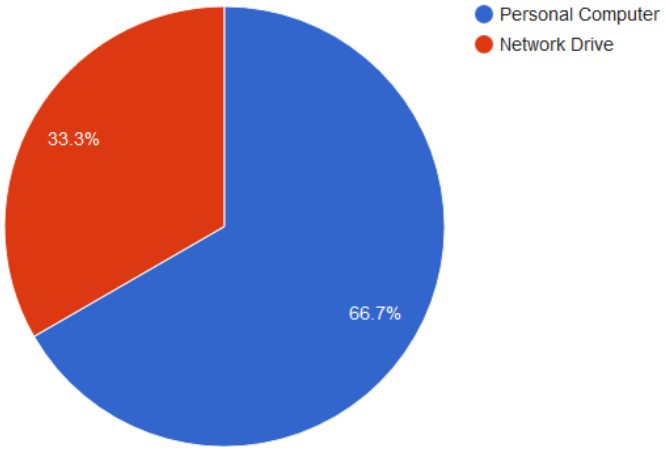


Standardized metadata, a set of data that describes and gives information about other data, is rarely used among those interviewed. Most participants use a "homegrown system" that works within their department or group of collaborators. The mostly commonly named form of metadata was a README file providing a description of the data. The responses reflect the interviewees focus on their individualized research, as opposed to the long-term preservation and usability of their data within the larger scientific community. An opportunity for educational outreach about discipline-specific metadata standards, as well as data curation efforts across the LAB, is evident in these responses.

> "Not really. Since I do not produce the data, I generally just use those that are associated with the data set I am using and manipulating. If I was to produce new 'data', like in the form of modeled output, I usually just describe this with the appropriate units."
>
> "We kind of have a homegrown metadata scheme: report numbers, causal analysis, a severity index (1-5) to triage level of damage. It could be better."
>
> "I don't even know what that means."
>
> "We capture all of the metadata in a README file. We don't necessarily use a standard or disciplinary metadata scheme."
>
> "Follow loose guidelines and use common terms. People always talk about standardized metadata in the field, but there is still no standard. I just create a readme file."
>
> "Sounds like more work."

Data Storage

Data storage includes all methods researchers use to store and back up data. Not surprisingly, the majority of interviewees store their data on their own personal computer and backup their data onto a network drive. At least 11 respondents spoke directly to the issue of storage space constraints. Several interviewees discussed the necessity of deleting "old" data off of servers to preserve space and cost. While several researchers frequently supplement network file storage with external hard drives and flash drives, a clear need for a long-term storage solution was expressed by the majority of researchers.



"If there was a better way to archive and access data it might be saved and that might mean less experimental data production, more mining of that archived data."

"Honestly, most experimental data is discarded because a very small percentage is used in analysis. Whatever is left over is currently stored of my hard drive."

"We save everything on different hard drives. We back everything up on M-Discs because they are cheap ($1) and there is no institutional solution. Honestly, we probably need a better solution. M-Discs are guaranteed for 100 years, but who knows if they will last that long."

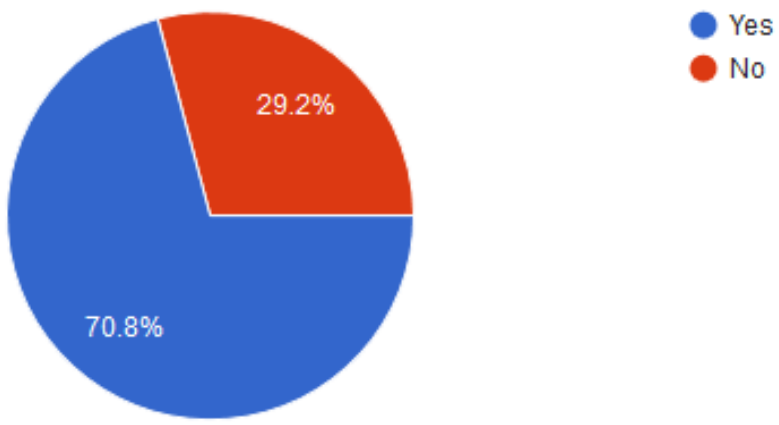
"My mentor and the PI on the grant is leaving at the end of the year and taking her funding with her. I will be converted, but I will work on the programmatic side of SIGMA. Not sure what will happen with all of that archived data."

"My own personal data is stored everywhere (flash drives, hard drive, floppy disks, magnetic tapes)."

"Data is mostly stored on everyone's individual computer. I did set up a drive on the shared drive for us to at least share images and files. There is no long-term storage solution."

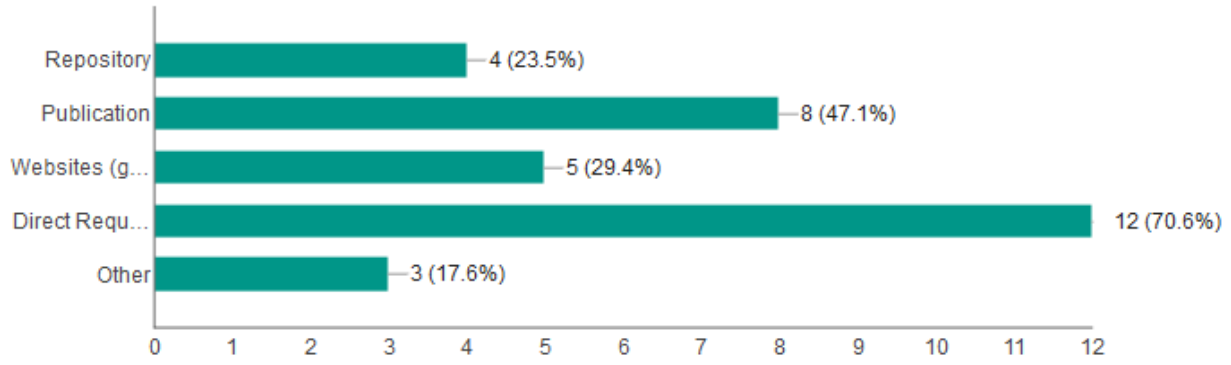
Data Sharing

A majority of interviewees share their data outside of the Lab. Because all of our interviewees work on non-classified projects, openness with the larger scientific community is a byproduct of the research process. Moreover, the mandate for environmental compliance to be publicly available is also reflected in this result. When interviewees were asked if they had any concerns about sharing data, the overwhelming majority of researchers had no issues, just wished to receive credit for their work.



The number of interviewees sharing their data in publications pointed to a graph or chart as evidence of data sharing. Of the 24 interviewees, only one researcher identified depositing a data set in a publisher-identified data repository.

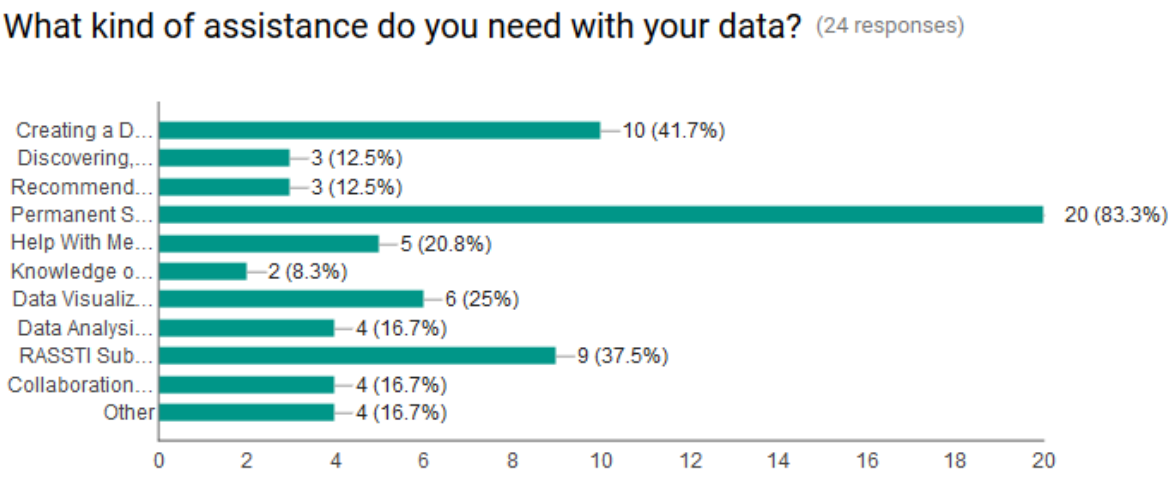
> "I am depositing my data in Figshare because I found a Nature article that said it was the best place to deposit your data (free, creates DOIs)."
>
> "No, our only concern is making deadlines due to compliance. Our data is all public. We share everything through the EMS web interface."
>
> "No, only to follow lab policy. I had a request for my research from Chinese researchers so I made sure to submit my materials through RASSTI."
>
> "No, understand that LANL owns all of our data. If they want it to be open, we make it open. Just want to make sure researchers/paper are acknowledgement /cited when data is used. A license of usage would be nice."
>
> "No, just credit. When you retire you start thinking about your life's work more."



A majority of researchers share their data with outside collaborators from direct requests from interested researchers. When asked how they share their data with outside collaborators, the majority of researchers use LANLTransfer to move their data. This commonly used protocol provides an area of exploration for use in RASSTI data set deposit workflow.

> "Most dissemination is based around specific requests from researchers, use LANL transfer, FTP to move data. A website where we could post/share data would be ideal."
>
> "I am interested in common platforms for publishing data or an easy mechanism to distribute (LANL firewalls make movement of data difficult). A platform like IETR from CERN would be useful."
>
> "Deposit all my data in NIST. Also, through RASSTI. I feel if I put more data out in the world it will be used. It's kind of an experiment."
>
> "We collaborate with Livermore and Sandia. We use LANLTransfer to send data to Sandia. Sometimes we use RASSTI; sometimes we don't."
>
> "We are currently working on a more elegant solution than LANLTransfer or shipping a hard drive physically (which is how it's done now)."
>
> "Honestly, when we get a request for our data we refer them back to the paper because our data output is complete. We very rarely, if ever, share more data than that. Once the paper is published, I wash my hands of it. For collaborators, we do share and received data using LANL Transfer."

Data Assistance



At the end of the interview, participants were asked what type of assistance they need in managing research data. Correlated with earlier responses, the three main points of need were:

1.) Creating data management plans

"Would be nice if LANL had 'default language' that I could just plug into the DMP (like storage and sharing polices)."

"It would be nice if there was some text about data sharing at the Lab that I could just plug into the DMP."

"It would be helpful to know how to draft a data management plan because I think those are becoming more common."

"I feel like most people take the 'ostrich' approach to mandates and DMPS - head in the sand will allow them to claim ignorance. It would be nice to have a clearer understanding of how LDRD proposal works with DMPTool."

2.) Permanent Storage

"A localized storage solution would be great. Wouldn't have to keep all my files on an office computer."

"It would be great if the Lab had a centralized repository where I could collaborate, store, and share my data both internally and externally. Something like Dropbox would even be helpful. I think Argonne has a wiki server that is accessible to outside researchers."

“Really, it would be nice to just know what options are available at LANL for storage and processing.”

““Much of this data is hosted on old websites that we maintain (scripted in PURL; accessible through FTP). In other words, old crusty crap.”

“A permanent storage solution, enhanced with metadata, would be ideal. Something like the Materials Data Facility at Argonne would be great.”

“We follow the clean water act policy of a 3-year retention policy, but our records still have a lot of redundancy. A local records management repository would be of great interest. Our Access database is reaching its limit of utility.”

3.) Submitting data through RASSTI

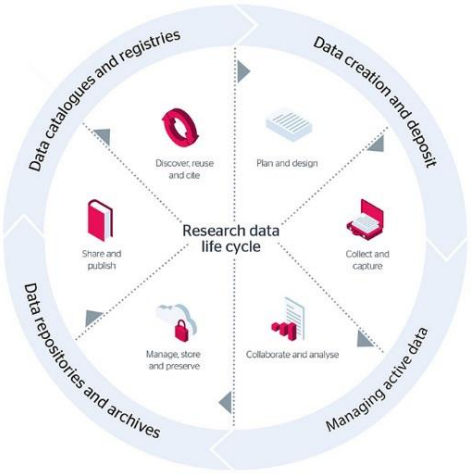
“It would be nice if RASSTI could handle data for release. I have a feeling that with the mandates coming down that more and more researchers will have to send their data through RASSTI.”

“RASSTI process is okay, but I've not had the greatest experience with SAFE1 folks. Data management sounds a lot like the old "quality assurance" initiative at the lab a few decades ago. It just gets in the way to doing the actual research.”

“Need RASSTI improvements.

“RASSTI is a pain.”

Conveniently, the three main areas of assistance identified by the Data Working Group align with the three most important components of the research data lifecycle, from planning to preservation to sharing.



Works Consulted

DataONE. (n.d.). Retrieved August 30, 2016, from https://www.dataone.org/

DCC. (n.d.). Digital Curation Centre | because good research needs good data. Retrieved September 1, 2016, from http://www.dcc.ac.uk/

FORCE 11. (2014, September 3). The FAIR Data Principles. Retrieved August 31, 2016, from https://www.force11.org/group/fairgroup/fairprinciples

Lake, S. (2013). *SPEC Kit 334: Research Data Management Services*.

Staff, P. R. G. (2016). *International Survey of Academic & Research Library Data Curation Practices, 2016-17*. Primary Research Group Inc.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLOS ONE*, *6*(6), e21101. http://doi.org/10.1371/journal.pone.0021101

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLOS ONE*, *10*(8), e0134826. http://doi.org/10.1371/journal.pone.0134826