Geoscience Digital Data Resource and Repository Service (GeoDaRRS) Workshop Report

Matthew S. Mayernik
Doug Schuster
Chung-Yi Hou
Greg Stossmeister

NCAR Technical Notes
NCAR/TN-552+PROC

National Center for
Atmospheric Research
P. O. Box 3000
Boulder, Colorado
80307-3000
www.ucar.edu

NCAR TECHNICAL NOTES

http://library.ucar.edu/research/publish-technote

The Technical Notes series provides an outlet for a variety of NCAR Manuscripts that contribute in specialized ways to the body of scientific knowledge but that are not yet at a point of a formal journal, monograph or book publication. Reports in this series are issued by the NCAR scientific divisions, serviced by OpenSky and operated through the NCAR Library. Designation symbols for the series include:

EDD – Engineering, Design, or Development Reports
Equipment descriptions, test results, instrumentation, and operating and maintenance manuals.

IA – Instructional Aids
Instruction manuals, bibliographies, film supplements, and other research or instructional aids.

PPR – Program Progress Reports
Field program reports, interim and working reports, survey reports, and plans for experiments.

PROC – Proceedings
Documentation or symposia, colloquia, conferences, workshops, and lectures. (Distribution maybe limited to attendees).

STR – Scientific and Technical Reports
Data compilations, theoretical and numerical investigations, and experimental results.

The National Center for Atmospheric Research (NCAR) is operated by the nonprofit University Corporation for Atmospheric Research (UCAR) under the sponsorship of the National Science Foundation. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

National Center for Atmospheric Research
P. O. Box 3000
Boulder, Colorado 80307-3000

NCAR/TN-552+PROC
NCAR Technical Note

2018 - 11

Geoscience Digital Data Resource and Repository Service (GeoDaRRS) Workshop Report

Matthew S. Mayernik
NCAR Library
National Center for Atmospheric Research, Boulder, CO
Doug Schuster
Computational & Information Systems Laboratory
National Center for Atmospheric Research, Boulder, CO
Chung-Yi Hou
Computational & Information Systems Laboratory
National Center for Atmospheric Research, Boulder, CO
Greg Stossmeister
Earth Observing Laboratory
National Center for Atmospheric Research, Boulder, CO

NCAR Laboratory
NCAR Division

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH
P. O. Box 3000
BOULDER, COLORADO 80307-3000
ISSN Print Edition 2153-2397
ISSN Electronic Edition 2153-2400

Geoscience Digital Data Resource and Repository Service (GeoDaRRS) Workshop Report

Contents

Acknowledgements

The National Science Foundation (NSF) provided the funding support for this workshop. We thank the steering committee members for their contributions to the workshop: Robert R. Downs, Danie Kinkade, Tran Nguyen, Mohan Ramamurthy, and Fuqing Zhang. We also thank Cecilia Banner and Elizabeth Faircloth of NCAR for administrative and logistical support. Thanks to Seth McGinnis and Jeff de La Beaujardière for providing detailed comments on this report.

1. Executive Summary

The National Science Foundation (NSF) funded Geoscience Digital Data Resource and Repository Service (GeoDaRRS) workshop was held from August 7-9, 2018 in Boulder, CO. The goal of the workshop was to help the NSF and data repositories better support geoscience researchers when they ask the question: “Where do I put my data?” Geoscience researchers are now being asked by funding agencies and scientific publishers to archive and cite data to support open access, but often struggle to understand and fulfill these requirements. The workshop brought together over sixty individuals from multiple stakeholder groups to discuss data management and archiving challenges and opportunities within the geosciences. The relevant stakeholder communities represented by the attendees included geoscience researchers, technology experts, scientific publishers, funders, and data repository personnel.

The workshop agenda was structured in four parts, with each part focusing on a different theme related to data management resources including: 1) What is the existing landscape and what gaps exist in that landscape for data producers and users, 2) What resources would you like to have and why, 3) What is realistic/doable with constraints, and 4) How do you move forward and act upon this discussion with your community? For each theme, a set of plenary speakers presented case studies, current projects, and personal perspectives. Additionally, workshop participants were asked to complete a pre-workshop survey, which included questions related to each theme. Following the plenary presentations, a set of breakout sessions occurred in which participants were led through guided discussions. These discussions allowed the participants to reflect back on the plenary presentations, pre-workshop survey results, and to address other related issues. The high-level insights from each breakout session were then presented back to the full workshop audience.

A number of recommendations were developed as a result of the workshop discussions. The recommendations were constructed to address challenges related to data management topics that repeatedly arose throughout the workshop. A summary of workshop recommendations by data management topic is provided below.

- **Long-term, Scalable Data Curation**
  - Long-term support for the data curation needs of the geoscience community is critical for providing truly open access to data. Several different, potentially combined, approaches to providing sustainable open access to data are possible, including: 1) augment existing geoscience data repositories to scale up their capacity, 2) identify non-specialized data repositories that fulfill open access objectives, 3) develop a data repository liaison service, and 4) create new data repository services.
- **Education and Training**
  - There is a need for programs to better support data management education within scientific, computing, information, and data disciplines.
  - Data management training and resources for researchers need to be improved, and better publicized.
- **Data Management Plans (DMPs)**
  - Grant proposal reviewers should 1) review data management planning according to multiple explicit criteria including sufficiency, resourcing, and execution, and 2) scrutinize this section as critically as all other sections of a proposal.
  - An efficient mechanism for grantees to update and comment on their DMPs during the annual reports would help improve accountability for the DMPs.
  - Data repositories need to be brought into the DMP conversations in the initial stages of the project planning process.
  - The geoscience community should foster a DMP tool ecosystem.
- **Funder & Publisher Policies**
  - All stakeholders should be clear on the core drivers and principles that motivate their data policies.
  - Coordination among all stakeholders is necessary to bring consistency to the data policy landscape.
  - Specific scientific research communities need to discuss and formalize data retention guidelines.
- **Strategic Partnerships**
  - Strategic partnerships across federal agencies could reduce costs through shared data storage and curation services.
  - Strategies need to be developed at the agency level to employ cloud computing and storage.
- **Legacy Data**
  - Researchers need clear paths to support curation and rescue of data collected via past projects.
- **Tools & Services**
  - All stakeholders should recognize the importance of open source software communities, and contribute to these efforts where possible.
  - Data repositories should investigate whether cost efficiencies can be gained by sharing data storage infrastructure.

For additional information, please find a detailed summary of challenges, recommendations, and desired outcomes organized by topic in Appendix VII.

2. Introduction

The open availability and wide accessibility of digital scientific resources, such as articles and datasets, is becoming the norm for 21st century science. The National Science Foundation (NSF) instituted a data management planning requirement in 2011, and many scientific journals in the geosciences have also recently implemented data archiving and citation policies (e.g. journals published by the American Geophysical Union and the American Meteorological Society). A 2013 U.S. White House Office of Science and Technology Policy (OSTP) memo emphasized the many benefits to science and society that could come from making scientific data and research information more accessible (Holdren, 2013). The OSTP memo recognized, however, that ensuring the open availability and reliable accessibility of articles, data, software, etc., would involve overcoming various challenges. To fulfill the promise of new open data initiatives, three key areas that need resolving are: 1) scientific resources (e.g. data and software) must be collected and documented properly, 2) repository services, including preservation and storage capabilities, must be maintained, supported, and improved over time, and 3) governance, including legal issues relating to copyright and resource ownership, must be established. Since the OSTP memo was issued, a number of federal agencies, including the NSF, have produced policy documents that describe processes to support the public access to data and other research results.

High-level policies and plans such as those noted above promote a strong culture of data sharing within the geosciences. Many disciplinary data facilities exist around the community to provide technical support and expertise for archiving data of particular kinds, or for particular projects. Within the geosciences, some research projects are well-supported by these existing data management and archiving facilities. However, many projects do not have the same level of facility support due to a combination of several factors, such as their smaller scale, funding limitations, topic scope that does not have a clear facility match, or uninformed data management practices. These projects typically manage data on an ad hoc basis without having long-term management and preservation procedures, beyond creating backup copies of data.

This report presents the outcomes of the Geoscience Digital Data Resource and Repository Service (GeoDaRRS) workshop. This workshop focused on establishing geoscience community expectations and requirements for digital data management support. Workshop attendees were asked to explore whether new services are needed to complement existing NSF-funded data facilities, particularly in the areas of: 1) data management planning support resources and 2) repository services for geoscience researchers who have data that do not fit in any existing repository. Functionally, new services would support NSF-funded researchers in meeting open access requirements set by the NSF and publishers for geosciences, thereby ensuring the availability of digital data for use and reuse in scientific research going forward.

The three-day workshop was held on August 7-9, 2018, hosted by the National Center for Atmospheric Research (NCAR) in Boulder, CO. It brought together over sixty individuals from multiple stakeholder groups to identify and outline requirements for repository services, and to discuss data management and archiving challenges and opportunities more broadly. The

relevant stakeholder communities represented by the attendees included geoscience researchers, technology experts, scientific publishers, funders, and data repository personnel.

3. Structure of the Workshop & Common Themes

A key goal of the GeoDaRRS workshop was to help NSF and data repositories to better support geoscience researchers when they ask the question: “Where do I put my data?” Geoscience researchers are now being asked by funding agencies and scientific publishers to archive and cite data to support open access, but often struggle to understand and fulfill these requirements. A pre-workshop survey (see results in Appendix VI) with questions related to data management requirements was administered to engage participants leading up to the workshop, and to help in motivating workshop discussions. The workshop agenda was structured in four parts, with each part focusing on a different theme related to data management resources.

- Theme #1: What is the existing landscape and what gaps exist in that landscape for data producers and users?
- Theme #2: What would you like to have and why?
- Theme #3: What is realistic/doable with constraints?
- Theme #4: How do you move forward and act upon this discussion with your community?

For each theme, a set of plenary speakers presented case studies, current projects, and personal perspectives. Following the plenary presentations, a set of breakout sessions occurred in which participants were led through guided discussions (See Appendix IV for breakout session discussion questions). These discussions allowed the participants to reflect back on the plenary presentations, pre-workshop survey results, and to address other related issues. The high-level insights from each breakout session were then presented back to the full workshop audience.

Below, the challenges and recommendations sections include key findings associated with each of the four above themes. These findings were pulled from the breakout discussions along with the insights from the plenary speakers. In the sections below, the term “data” refers to datasets and data collections, and does not encompass other kinds of scientific resources like software and model code. Recommendations or findings specific to software or other resources are called out explicitly. For detailed notes organized by theme, please see Appendix V.

4. Challenges

The GeoDaRRS presentations and discussions touched on a number of tensions related to data sharing that have existed within scientific institutions for hundreds of years. This section describes some of these challenges and contextualizes them via prior literature.

The first set of challenges highlighted here relates to the mixed incentives that researchers face with regards to sharing data. Researchers must navigate norms such as transparency and reproducibility, while ensuring their own competitive advantage by maximizing their access to novel data, tools, or knowledge (Merton, 1942/1973; Mitroff, 1974). In today's rapidly evolving policy climate, researchers' data practices are expected to be robust and structured and conducive to data sharing. The ubiquity of the internet amplifies these expectations, as it provides a seemingly universal data distribution mechanism (Agre, 2002). ***Many impediments to data sharing still exist, however. The most salient disincentives to data sharing are well known, such as the potential for researchers to be "scooped" on important scientific findings, the possibility for data to be misunderstood by secondary users, the lack of clear reward structures for data sharing, and the time and effort required to clean, document, and provide access to data (Arzberger et al, 2004; Borgman, 2012).*** As summarized in a 1985 National Academies report titled "Sharing Research Data": "Although it serves science for researchers to share their data and permit reanalysis and replication, it is often not in their interest to do so" (Fienberg, Martin, & Straf, 1985, pg. 17). These disincentives to share data have been repeatedly confirmed by empirical studies (cf. Campbell et al, 2002; Tenopir et al, 2011; Federer et al, 2015).

Many of these disincentives were discussed during the GeoDaRRS presentations and breakout discussions. ***Researchers face many practical constraints and significant demands on their time and expertise.*** Time is a limitation on data management for many researchers. It can take a significant amount of time to process data and to produce complete and quality metadata, all of which are necessary to get data into a repository. These time limitations can result in "throw it over the fence" workflows between researchers and data repositories. ***Ensuring credit for authors of datasets with many contributors is another widely noted problem.*** Datasets used to write scientific papers are often composites that contain data from many individual instruments, collections, and models. Group authorships are problematic, and no well-established norms exist around how to acknowledge and reward the diverse contributions to data work (Brand et al, 2015; Hou & Mayernik, 2016). Reducing this problem is important. Without clear demonstrations of direct benefits gained by including data management steps throughout the full research lifecycle (from research proposal to research output), researchers will inconsistently prioritize data management.

Another significant challenge for data archiving and sharing discussed during the GeoDaRRS sessions is that data do not exist in isolation. Datasets are produced as one component within the larger research process that also includes research methods, instruments, software, hardware, and documentation. In addition, there are many different kinds of data, on a spectrum from purely observations to purely model simulation output, and from physical to chemical to biological to socio-economic, with many types of combined and derived data in between. In many collaborative research settings, responsibilities for keeping track of all of these various components are distributed and vaguely defined (Wallis & Borgman, 2011; Leonelli, 2016). ***Researchers lack clear guidance or recommendations on what components and granularities of their research process should be archived, and how to ensure the relationships among them are not lost.***

The GeoDaRRS workshop also featured much discussion about the constraints and challenges that data repositories face. While numerous data facilities exist for the archiving of geoscience data, ***these existing data facilities are limited in what they can accommodate.*** Repositories face practical constraints related to technical, human, and financial factors. From a technical perspective, data storage requirements continue to grow, and repositories must continually refresh their software and hardware infrastructures as technologies evolve. From a human perspective, repositories build staff expertise to support specific kinds of data and data systems. Providing a high level of curation support for very diverse data can be difficult with a limited staff.

These technical and human factors are of course related to financial limitations that repositories face. Many geoscience data repositories are funded via short term grants or cooperative agreements that are subject to competitive bidding on regular intervals (often every five years). Repositories in some cases add a mix of grant funds, fees for services, and community memberships to buttress operating costs (Mayernik et al, 2012). But as noted by multiple GeoDaRRS plenary speakers, ***planning, expanding, and sequencing data repository operations are significant challenges without commitments of continuity from funders.*** Beyond these challenges in sustainability and continuity, repositories face increasing and uncertain costs associated with data storage and supporting career paths for data professionals. GeoDaRRS attendees noted that there is a need to have a conversation about funding agencies' roles in paying for storage and archiving costs, and developing career path opportunities for data professionals.

Finally, as the above discussion indicates, ***understanding how to manage, curate, and preserve data optimally is itself an area of science.*** Active topics of current research within the computer, information, and data sciences include (but are not limited to) how to build sustainable and effective data infrastructures, how to integrate metadata into big data and data science workflows, and how to characterize social and institutional challenges relating to data curation (Fox & Hendler, 2014; Borgman, 2015; Greenberg, 2017).

5. Recommendations

Recommendations from the workshop are organized according to topics that were repeatedly highlighted in the workshop plenary presentations and breakout discussions (For detailed notes, see Appendix V). These include:

- 5.1 Long-term Data Curation
- 5.2 Education and Training
- 5.3 Data Management Plans
- 5.4 Funder & Publisher Policies
- 5.5 Strategic Partnerships
- 5.6 Legacy Data
- 5.7 Tools & Services

These recommendations are designed to support the data management needs of geoscience researchers, and to generally enable data products to be more readily available, understandable, and preservable for the benefit of scientific research. The recommendations are not exhaustive or ordered in any way, and progress on any topic would be beneficial. A summary of Workshop Recommendations by Topic is provided in Appendix VII.

5.1 Long-term, Scalable Data Curation

As mentioned above, researchers today face many requirements to cite and archive data from both publishers and funders alike. Open Data requirements (Holdren, 2013) pose an increasing burden on researchers and repositories. Both groups need assistance in meeting this challenge. In particular, some participants in fact voiced concerns that they do not have places where they could archive their data for public access, and where the data are also expertly curated. This perspective for example was expressed for air quality data, especially those that are associated with health and welfare. Likewise, multiple participants who conduct hydrological and/or atmospheric modeling expressed that there is a need for a resource where data products can be handed off once a grant is complete. A particular issue noted was the difficulty in archiving and supporting software and data products after short-term grants end. Projects that generate large data volumes, involve international teams, and/or involve interdisciplinary research face additional challenges. At present, individuals dealing with these challenges develop temporary, ad hoc solutions to manage these data issues, including putting data on cloud infrastructures, local research group servers, and university-operated servers.

5.1.1 Recommendation: ***Long-term support for the data curation needs of the geoscience community is critical for providing truly open access to data.*** Several different approaches to providing sustainable open access to data are possible. Further investigation will be needed to understand the relative merits and drawbacks of each approach for current and future needs. Any decisions about these approaches should be based on input from data repositories and research grantees. Ultimately, guidance regarding the use of these approaches may involve some combination of all four approaches, perhaps in a distributed model where organizations share expertise and/or infrastructure.

1. ***Augment existing geoscience data repositories to scale up their capacity.*** As noted above in the Challenges section, the existing data repositories that focus on geoscience data are constrained in their ability to scale up to meet new data archiving demands. Finding resources to scale these repositories up, however, could be an efficient way to build on existing capacities and expertise within the data repository landscape. Some repositories might be most constrained by technological limitations, e.g. data storage space or data delivery technologies. Other repositories might be more constrained by staffing limitations, e.g. lack of expertise in working with data with new kinds of data formats or topic areas. Direct input from repositories will be necessary to understand what additional scale is possible if they are provided with increases in resourcing.

2. ***Identify non-specialized data repositories that fulfill open access objectives.***
Geoscience-focused repositories represent only part of the data repository landscape. Many libraries within research universities offer data curation and archiving services of some kind (Cox et al 2017; Coates et al, 2018). University library services vary in detail and cost, and not all universities offer such a service. Where they exist, however, they provide a local source of expertise and infrastructure, and they can work with researchers across academic disciplines. Another type of non-specialized data repository are general-purpose repositories such as Zenodo, Dryad, Figshare, and the Open Science Framework (OSF). These are web-based services that provide places to post data online and ensure public access. Their costs to the user are generally low or free for small datasets (10-20 Gb), with potentially increasing costs for increasingly larger data volumes. These services, however, in some cases provide minimal or no data curation support (Zenodo, Figshare, and OSF), leading to significant questions about data quality, persistence, and consistency. The use of such services should be coupled with data curation consulting to ensure accountability, transparency, and usability of the data (Mayernik, 2017). Finally, existing efforts, such as those supported by the NSF Office of Advanced Cyberinfrastructure (OAC), should be better promoted, as many of the products developed through these programs address the data curation support needs for a variety of scientific communities.

3. ***Develop a data repository liaison service.***
The data repository landscape described in points #1 & 2 above is large and evolving continuously. Researchers cannot be expected to understand or evaluate all of the options that may be available to them. New tools have potential to help researchers identify appropriate repositories, but such tools are at best in prototype phase (such as https://repositoryfinder.test.datacite.org/), and will inevitably struggle to remain current and accurate. Creating a liaison service may have significant benefit in assisting researchers in finding a repository. This liaison service might take the form of a "help desk" or a "network of experts," but the role of the service would be to provide researchers with direct advice on finding an appropriate repository, and potentially could also assist researchers with the process of data submission. Such a service would need to be scoped and coordinated, but it could build on existing networks of geoscience data professionals, publisher services, and university data librarians, including the Earth Science Information Partners (ESIP) and the Research Data Access and Preservation (RDAP) networks.

4. ***Create new data repository services***
The GeoDaRRS presentations and discussions identified a number of characteristics of effective data repositories that should be emphasized if new data repository services are to be built. These can be considered to be requirements for any new data repository.
    - Build from community-established principles, such as the Findable, Accessible, Interoperable, Reusable (FAIR) guidelines (Wilkinson et al, 2016), and from internationally-recognized "trusted repository" certifications, such as the CoreTrustSeal (https://www.coretrustseal.org/).

- Emphasize curation - It is one thing to store data, and it is another thing to curate data to enable understanding and use. As one workshop presenter stated: “Do not just build a bit bucket.”
- Leverage and build upon relevant products being developed through existing grant programs. Examples include programs supported by the NSF OAC (e.g. DataNet/Data Infrastructure Building Blocks (DIBBs), CyberInfrastructure for Sustained Scientific Innovation (CSSI), EarthCube, etc…).
- Adopt common ways of serving data via machine-readable services in collaboration with concurrent community efforts (e.g. National Institutes of Health Data Commons and NSF DIBBs funded projects) - Researchers would love to be able to write three lines of code to pull data into processing pipelines, but they do not want to have 100 different scripts to access data from different repositories.
- Account for Intellectual Property concerns - Enable limited embargoes on data access, “private” repository work spaces, and appropriate data licensing frameworks.
- Use metadata schemas that are well defined, yet flexible enough to deal with unique aspects of scientific research (obscure instrument types, etc)
- Archive data in formats that are aligned with community standards (e.g. netCDF that conforms to the CF, or Climate and Forecast, metadata conventions)
- Use persistent identifiers to enable web-based identification and location of data.
- Provide scalable data storage and co-located scalable compute/analysis capabilities to service large volume datasets.

Some additional considerations for any new repository touch on financial and social issues:

- Data repositories need long-term, sustained funding that is not project specific.
    - A big cost for data repositories is in data ingestion. There is typically an initial burst of effort when data arrives. After ingestion, storage and maintenance costs are more salient, although value-added user access/compute costs can be unpredictable.
    - Coordinated and consistent community engagement is critical when developing new services/repositories. This requires time, staff, and additional resources to effectively engage the appropriate stakeholders. Frequently engaging stakeholders when developing a new service, similar to the Agile software development and usability testing model, is challenging, but essential. It is important not to assume stakeholder needs.

5.2 Education & Training

GeoDaRRS discussions emphasized the need to integrate data management and archiving practices into education and training curricula. Many problems that we encounter today when dealing with data result from people not being aware of data and data management practices in the past. Data management education and training should be integrated into undergraduate and

graduate curricula, to enable knowledge flows from senior researchers to students, and vice versa. For example, graduate students learn about what is important in their research area through reading published papers. Seeing data archiving and sharing approaches described in published papers is therefore an important way to spread desired practices. Conversely, because graduate students perform much of the day-to-day work in academic research, they have a significant role in bringing new data management approaches into research teams, thereby effecting change in how entire labs or teams operate.

A second issue related to education and training is that few formal educational programs have been established for developing professionals who are knowledgeable in scientific computing and data. Scientific institutions have difficulty hiring people with the skill sets required to work on cloud-based systems, for example.

5.2.1 Recommendation: ***There is a need for programs to better support data management education within scientific, computing, information, and data disciplines.*** These programs will need to be designed appropriately to engage and involve scientists who face significant demands on their time. Specific types of programs mentioned during the GeoDaRRS workshop include:

- Targeted internships, for example, supporting science student internships to work in a data repository. These types of internships could teach students about data management principles and practices via direct work experience.
- Workshops where people "do" rather than just listen (e.g. hackathons). Bring scientists and data repository staff together to better define common approaches and best practices for data management.

5.2.2 Recommendation: ***Data management training and resources for researchers need to be improved and better publicized.***

- For scientists, training on data management best practices and writing good data management plans is essential to engender thinking about data management from the beginning of the project.
    - The Data Management Training Clearinghouse (http://dmtclearinghouse.esipfed.org/) provides a registry of data management training resources.

5.3 Data Management Plans (DMPs)

The NSF instituted a data management planning requirement in 2011. This requirement is still the main policy mechanism used by the NSF to promote and engender more robust data management within the sciences. Much discussion during the GeoDaRRS workshop focused on this DMP requirement, and how to make it more effective. A general point made repeatedly is that data management plans need to be scrutinized by reviewers just as critically as the rest of the proposal to ensure they are well thought out and effective. Researchers also need ways to ask for help with data management without being penalized (in actuality or in perception) for needing help. In particular, principal investigators would gain significant benefit from receiving

feedback from reviewers and program officers on their DMPs, and from knowing how the DMPs are evaluated.

Many different national and international groups have developed recommendations and tools for improving the process of writing, evaluating, and executing data management plans. A central point of these recommendations and the DMP discussions at the GeoDaRRS workshop is that the DMP document itself should be seen as the beginning of the process of data management, not the end point. As such, guidelines/instructions for proposal and publication reviewers should emphasize that DMP review involves looking at the DMP in the context of the full proposal and over the full grant timeline. Further, finding ways to connect data repositories into the DMP process could significantly increase the quality of the DMPs and the ultimate data management outcomes. Finally, a variety of DMP tools offer the possibility of increasing the robustness of DMPs during the proposal process and afterward.

5.3.1 Recommendation: *Grant proposal reviewers should 1) review data management planning according to multiple explicit criteria including sufficiency, resourcing, and execution plans, and 2) scrutinize this section as critically as all other sections of a proposal:*

1. Sufficiency - does the DMP itself address the important components?
2. Resourcing - is data work written into the grant budget appropriately?
3. Execution - does the DMP and/or proposal narrative describe how the data management tasks will be achieved, e.g. who will be responsible for the tasks, how these tasks will be sequenced with other project work, and how success will be evaluated?

5.3.2 Recommendation: *An efficient mechanism for grantees to update and comment on their DMPs during the annual reports would help improve accountability for the DMPs.* Funding agencies already ask grantees to report on a project's progress on a yearly basis. These yearly report submissions could provide a periodic way for grantees to report on their data management activities. Including a discussion of what data management activities are happening concurrently would properly recognize the importance of the DMP to the project. However, such reporting requirements should not put too much time burden on researchers.

5.3.3 Recommendation: *Data repositories need to be brought into the DMP conversations in the initial stages of the project planning process.* When researchers know specifically where they plan to deposit data, they are much more likely to follow through with their plan. Data repositories can also provide specific examples and details that can inform project data management tasks and costs.

5.3.4 Recommendation: *The geoscience community should foster a DMP tool ecosystem.* This support might take two forms:

- Proposal writers should be encouraged to use existing DMP tools, in particular, the DMPTool (https://dmptool.org/) developed by the California Digital Library. This tool has existed for number of years and provides templates that meet DMP requirements of numerous funders.

- Develop and implement new tools to support DMP development.
    - DMPs are written and submitted as text. Stakeholders, such as funders or supporting institutions, might work to transition free-text DMPs to a form-based approach where DMP information is more highly structured and specified. This would also facilitate enabling DMP information to flow to other implicated stakeholders, such as administrative and data repository staff.
    - DMPs are currently static. Research projects, however, are highly iterative and may evolve significantly during the life of a grant. Stakeholders might work to transition toward active data management plans which can be iteratively reviewed and adjusted.
    - Both of the above points suggest that there would be utility in developing machine-actionable DMPs. The idea behind machine-actionable DMPs is that they are able to be automatically generated and shared with collaborators and funders (Miksa et al, 2018). A number of international groups are working on standards and procedures for machine-actionable DMPs (RDA, 2018a,b). The NSF has also funded multiple projects to investigate how to convert DMPs into dynamic data feeds ("Making Data Management Plans Actionable," https://nsf.gov/awardsearch/showAward?AWD_ID=1745675 ; "Supporting Public Access to Supplemental Scholarly Products Generated from Grant Funded Research**,"** https://nsf.gov/awardsearch/showAward?AWD_ID=1649703).

5.4 Funder & Publisher Policies

Researchers face a variety of data management requirements and recommendations. Inconsistent data management implementation guidelines across NSF programs and directorates adds to the challenge of meeting compliance requirements. Researchers who work with different funders face even more complexity in dealing with variable data management recommendations and requirements (Kriesberg et al, 2017).

In addition to inconsistent funder policies and guidance, researchers encounter significant variability in publisher data policies. Journal policies provide key leverage points in motivating and changing the data archiving and citation practices of scientific researchers (Kim & Stanton, 2016; Coulture et al, 2018), but inconsistency in journal policies (and their implementations) can confound researchers' attempts to publish their papers and archive their data. This is particularly important as some funders' data policies direct researchers to follow publishers' policies related to important data issues. One example is the NSF's 2016 public access plan (NSF, 2016) paragraph regarding "data deposit and citation":

> Data that underlie the findings reported in a journal article or conference paper should be deposited in accordance with the policies of the publication and according to the procedures laid out in the DMP included in the proposal that led to the award on which the research is based. (pg. 6)

If funders are deferring to publishers' data policies, it is critical for publishers and funders to be in sync with regards to data management requirements. AGU's "Enabling FAIR Data" project, funded by the John and Laura Arnold Foundation, is working toward harmonization of publishers' data policies (Stall et al, 2017). Continuing to move toward coordination among publishers and funders will streamline the compliance process for researchers.

Beyond the core GeoDaRRS question of "where do I put my data?", the key questions that challenge researchers with regards to meeting policy requirements are "what to save?" and "for how long?" Guidance on these topics is particularly vague and inconsistent across funders and publishers. While there is a need to enable flexibility for different types of data, such as model output vs. observational data, more detailed guidance on these questions is needed. Modelers who attended the GeoDaRRS workshop expressed significant confusion on these questions. Part of the challenge for researchers is that funders' data policies are often unclear about the motivations for their data policies. For example, "enabling data sharing" and "enabling results from published research to be reproduced" require different implementation approaches. Achieving one of these two goals does not necessarily mean achieving the other.

5.4.1 Recommendation: ***All stakeholders should be clear on the core drivers and principles that motivate their data policies.*** For example, ensuring reproducibility may involve different work than ensuring data accessibility. Roles and responsibilities associated with these drivers and principles should likewise be defined clearly. More clarity on these points should help with consistency in data policy implementations across stakeholders.

5.4.2 Recommendation: ***Coordination among all stakeholders is necessary to bring consistency to the data policy landscape.*** By contributing efforts towards data policy coordination, stakeholders can demonstrate their commitment to bring about positive change in the data policy landscape.

5.4.3 Recommendation: ***Specific scientific research communities need to discuss and formalize data retention guidelines.*** This was identified during the GeoDaRRS workshop as a particular need for the atmospheric and hydrological modeling communities. These guidelines might be developed via focused workshops, town hall meetings, or professional association working groups.

5.5 Strategic Partnerships

A repeated point of discussion during the GeoDaRRS workshop related to the need for organizations to develop partnerships to reduce duplication of efforts (and spending), and to facilitate shared solutions to common problems. For example, the NSF EarthCube program successfully worked with Google to support geoscience data facilities in being indexed by the new Google Data Search, https://www.earthcube.org/group/project-418. Other potential partnerships might focus on general challenges related to data storage and curation, or focus on more specific goals, such as developing better ways to serving streaming video data online.

5.5.1 Recommendation: ***Strategic partnerships across federal agencies could reduce costs through shared data storage and curation services.*** Federal agencies with geoscience foci, particularly the NSF, USGS, NOAA, the Department of Energy, and NASA, operate data centers as part of fulfilling their agencies' congressionally-mandated missions. These agencies have experience and technical tools associated with large-scale and small-scale geoscience data. For example, there may be opportunities to develop partnerships in which data funded by the NSF are archived in these federal data centers or vice versa. In this case, the NSF role would be to serve as an orchestrator of data archiving relationships, not a direct operator of data storage facilities. At minimum, the various agencies could work together to discuss common data infrastructure needs, for instance, to facilitate potential cost-sharing for co-located storage and cloud computing services. As an outside example, a group of university libraries have formed a network to develop strategic partnerships for data curation resource and expertise sharing (https://sites.google.com/site/datacurationnetwork/). These network models have potential to leverage distributed resources, and reduce duplication of efforts across organizations.

5.5.2 Recommendation: ***Strategies need to be developed at the agency level to employ cloud computing and storage.*** Cloud computing providers are part of the future for scientific data management, storage, and preservation. Some NSF-funded projects are leveraging or building cloud services, including projects funded by the NSF EarthCube program, Pangeo (http://pangeo.io/) and GeoSciCloud (https://www.nsf.gov/awardsearch/showAward?AWD_ID=1639709), and the NSF-funded Jetstream cloud infrastructure (https://jetstream-cloud.org/), but these projects are not coordinated or systematic for NSF as whole. Cloud storage and computing do not solve all of the issues that geoscientists and geodata repositories face. Less popular datasets may be deleted over time. Cloud users need to be mindful of matching cloud capabilities and costs to the dataset user community needs. Long term viability and costs of using the cloud are still an open question. Some agencies have had significant success with cloud-based strategies, such as the NOAA Big Data Project (Ansari et al, 2018), in which select NOAA operational datasets were made available via a combination of cloud storage providers. The NOAA data that were put into the cloud saw a significant increase in use.

5.6 Legacy Data

Many researchers currently store data collected via past projects. When the grants used to collect these legacy data expire, researchers have few options for data storage or archiving. GeoDaRRS workshop discussions generally emphasized that legacy data issues were important, but difficult to scope or address comprehensively. Attendees recommended that the geoscience repositories focus the majority of resources on present and future projects, while recovering legacy data as resources allow. New research may also drive legacy data activities, if new research questions or analytical techniques stimulate interest in particular legacy datasets. In general, legacy efforts should be supported according to science demands.

For observational data, model data, experiment data, and physical samples, legacy data issues have to be dealt with in different ways. Observational data, including physical samples, must be preserved indefinitely because they cannot be re-collected. Model and experiment produced data in some cases can be regenerated, or the models or experiments can be re-run to generate comparable data. Re-running models or experiments, however, can be time consuming and difficult. For example, preserving initial conditions, computational environments, and input data for atmospheric models is non-trivial while many of the legacy models or software may no longer be compilable or executable in new computer software or architecture.

5.6.1 Recommendation: ***Researchers need clear paths to support curation and rescue of data collected via past projects.*** These might include***:***

- Small grants to support researchers and/or data professionals to quality check, document, and reformat existing data and deposit them into a repository.
- Coordinated legacy data initiatives that partner with available data repositories. For example, an environmental data repository represented at the GeoDaRRS meeting put a call to its community for help in upgrading legacy data. Their call noted that if people were willing to upgrade their legacy data, the repository would work with them to serve up the data. This was effective at getting participation, and reducing the legacy data problem within their community.

5.7 Tools & Services

The GeoDaRRS workshop featured extensive discussions of tools and services for managing and analyzing data. Some of these topics are included in the recommendations above, as in the discussion of tools for DMPs and strategic partnerships at the agency level. Two other issues related to tools and services are described here: the importance of open source software communities, and the need to look at alternative models for data storage.

5.7.1 Recommendation: ***All stakeholders should recognize the importance of open source software communities, and contribute to these efforts where possible.*** Scientists push the boundaries of existing tools, often customizing or extending tools in unique ways. “Off the shelf” software is thus not always sufficient as a data analysis solution. Open source software has become instrumental in providing scientists with high-quality data analysis tools, along with communities of practice on how to best extend and contribute back to the same code bases. These communities are critical to the effectiveness of the open source software ecosystem.

5.7.2 Recommendation: ***Data repositories should investigate whether cost efficiencies can be gained by sharing data storage infrastructure.*** This could involve conceptual (and potentially financial) separation between the data storage function and the data curation function that repositories provide. This could also involve developing ways to adjust data storage tiers to accommodate variable costs in accordance with usage, e.g. allocate cheaper data storage options for data that rarely get used. The workshop discussions emphasized that data storage costs continue to be an ongoing challenge. Despite the decrease in storage costs on a per-unit

basis, total storage costs continue to increase as the amount of data generated through computational models and new observational instruments and sensors also increases. Workshop attendees discussed the need to investigate new models for sharing costs and leveraging Cloud technologies, as mentioned above. Some of these discussions about infrastructure sharing have already begun within the Council of Data Facilities, which is organized via the NSF EarthCube initiative: https://www.earthcube.org/group/council-data-facilities.

6. Conclusion

A number of data management challenges and potential pathways forward were discussed during the workshop. The recommendations outlined in this report are intended to provide concrete steps on how stakeholders can move forward and work to address these challenges. In addition to these recommendations, it was agreed upon that changes in the broader research culture will be needed to enhance the value of data management activities in research workflows. Progress in changing the research culture will require champions in disciplinary research communities to amplify the message and efforts of data professionals, and pass along their knowledge to colleagues. Finally, special sessions in disciplinary conferences and meetings to discuss data management related issues could provide an additional and valuable mechanism for community outreach, and to support culture change.

7. Bibliography

Agre, P.E. (2002). Real-time politics: The internet and the political process. *The Information Society*, 18(5): 311-331. https://doi.org/10.1080/01972240290075174

Ansari, S., Del Greco, S., Kearns, E., Brown, O., Wilkins, S., Ramamurthy, M., et al. (2018). Unlocking the potential of NEXRAD data through NOAA's Big Data Partnership. *Bulletin of the American Meteorological Society*, 99(1): 189-204. https://doi.org/10.1175/bams-d-16-0021.1

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., et al. (2004). Science and government: an international framework to promote access to data. *Science*, 303(5665): 1777-1778. https://doi.org/10.1126/science.1095958

Borgman, C.L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6): 1059-1078. https://doi.org/10.1002/asi.22634

Borgman, C.L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: MIT Press.

Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing*, 28(2): 151-155. https://doi.org/10.1087/20150211

Campbell, E.G., Clarridge, B.R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N.A. & Blumenthal, D. (2002). Data withholding in academic genetics: evidence from a national survey. *Journal of the American Medical Association*, 287(4): 473-480. https://doi.org10.1001/jama.287.4.473

Coates, H. L., Carlson, J., Clement, R., Henderson, M., Johnston, L. R., & Shorish, Y. (2018). How are we measuring up? Evaluating research data services in academic libraries. *Journal of Librarianship and Scholarly Communication*, 6(1). https://doi.org/10.7710/2162-3309.2226

Couture, J.L., Blake, R.E., McDonald, G., & Ward, C.L. (2018). A funder-imposed data publication requirement seldom inspired data sharing. *PLoS ONE*, 13(7): e0199789. https://doi.org/10.1371/journal.pone.0199789

Cox, A.M., Kennan, M.A., Lyon, L. & Pinfield, S. (2017). Developments in research data management in academic libraries: Towards an understanding of research data service maturity. *Journal of the Association for Information Science and Technology*, 68(9): 2182-2200. https://doi.org/10.1002/asi.23781

Federer L.M., et al. (2015) Biomedical data sharing and reuse: Attitudes and practices of clinical and scientific research staff. *PLoS ONE*, 10(6): e0129506. https://doi.org/10.1371/journal.pone.0129506

Fienberg, S.E., Martin, M.E., & Straf, M.L. (Eds). (1985). Part I: Report of the Committee on National Statistics. In *Sharing Research Data*. Washington, D.C.: National Academy Press. http://www.nap.edu/catalog/2033/sharing-research-data

Fox, P. & Hendler, J. (2014). The science of data science. *Big Data*, 2(2): 68-70. https://doi.org/10.1089/big.2014.0011

Greenberg, J. (2017). Big metadata, smart metadata, and metadata capital: Toward greater synergy between data science and metadata. *Journal of Data and Information Science*, 2(3): 19-36. https://doi.org/10.1515/jdis-2017-0012

Holdren, J.P. (2013). *Increasing Access to the Results of Federally Funded Scientific Research.* U.S. Office of Science and Technology Policy: Wash. D.C. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Hou, C.-Y. & Mayernik, M. (2016). Recognizing the diversity of contributions: A case study for framing attribution and acknowledgement for scientific data. *International Journal of Digital Curation*, 11(1): 33-52. https://doi.org/10.2218/ijdc.v11i1.357

Kim, Y. & Stanton, J.M. (2016). Institutional and individual factors affecting scientists' data sharing behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology*, 67(4): 776-799. https://doi.org/10.1002/asi.23424

Kriesberg, A., Huller, K., Punzalan, R., & Parr, C. (2017). An analysis of federal policy on public access to scientific research data. *Data Science Journal*, 16: paper 27. http://doi.org/10.5334/dsj-2017-027

Leonelli, S. (2016). Locating ethics in data science: responsibility and accountability in global and distributed knowledge production systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083). https://doi.org/10.1098/rsta.2016.0122

Mayernik, M.S. (2017). Open data: Accountability and transparency. *Big Data & Society*, 4(2). https://doi.org/10.1177/2053951717718853

Mayernik, M.S., Choudhury, G.S., DiLauro, T., Metsger, E., Pralle, B., Rippin, M., & Duerr, R. (2012). The Data Conservancy Instance: Infrastructure and organizational services for research data curation. *D-Lib Magazine*, 18(9/10). https://doi.org/10.1045/september2012-mayernik

Merton, R.K. (1942/1973). The normative structure of science. In Merton, R.K., *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago: University of Chicago Press.

Miksa, T., Simms, S., Mietchen, D., & Jones, S. (2018). Ten simple rules for machine-actionable data management plans (preprint version). http://doi.org/10.5281/zenodo.1172673

Mitroff, I.I. (1974). Norms and counter-norms in a select group of the Apollo moon scientists: A case study of the ambivalence of scientists. *American Sociological Review*, 39(4): 579-595. https://doi.org/10.2307/2094423

National Science Foundation (NSF). (2015). *NSF's Public Access Plan: Today's Data, Tomorrow's Discoveries. Increasing Access to the Results of Research Funded by the National Science Foundation*. NSF Report 15-52. https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf15052

Research Data Alliance (RDA). (2018a). *DMP Common Standards Working Group.* https://www.rd-alliance.org/groups/dmp-common-standards-wg

Research Data Alliance (RDA). (2018b). *Exposing Data Management Plans Working Group.* https://www.rd-alliance.org/groups/exposing-data-management-plans-wg

Stall, S., Robinson, E., Wyborn, L., Yarmey, L., Parsons, M., Lehnert, K., et al. (2017). Enabling FAIR data across the Earth and space sciences. *Eos*, 98. https://doi.org/10.1029/2017EO088425

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, 6(6): e21101. https://doi.org/10.1371/journal.pone.0021101

Wallis, J.C. & Borgman, C.L. (2011). Who is responsible for data? An exploratory study of data authorship, ownership, and responsibility. *Proceedings of the American Society for Information Science and Technology*, 48(1): 1-10. https://doi.org/10.1002/meet.2011.14504801188

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3: 160018. https://doi.org/10.1038/sdata.2016.18

8. Appendix I - List of Acronyms

AGU - American Geophysical Union
FAIR - Findable Accessible Interoperable Reusable
GeoDaRRS - Geoscience Digital Data Resource and Repository Service
DMP - Data Management Plan
ESIP - Earth Science Information Partners
NASA - National Aeronautics and Space Administration
NCAR - National Center for Atmospheric Research
UCAR - University Corporation for Atmospheric Research
NSF - National Science Foundation
NOAA - National Oceanic and Atmospheric Administration
OSF - Open Science Framework
OSTP - White House Office of Science & Technology Policy
RDAP - Research Data Access and Preservation
USGS - United States Geological Survey
OAC - Office of Advanced Cyberinfrastructure
DIBBS - Data Infrastructure Building Blocks
CSSI - CyberInfrastructure for Sustained Scientific Innovation

9. Appendix II - Workshop Participants and Steering Committee Members

The full attendee list can be accessed online at:
https://www2.cisl.ucar.edu/events/workshops/geodarrs-workshop/2018/participants-list

On-site Participants: Deb Agarwal, Franco Biondi, Jessica Blois, Ken Bowman, Sarah Callaghan, Laura Condon, Tom Cram, Linda Cully, Ethan Davis, Jeff de La Beaujardière, Robert R Downs, Ryan Frazer, Michael Friedman, Joe Hamman, Fred Harris, Dave Hart, Sophie Hou, Leslie Hsu, Cesunica Ivey, Nick Jarboe, Anke Kamrath, Danie Kinkade, Rebecca Koskela, Madison Langseth, Kerstin Lehnert, Reed Maxwell, Matt Mayernik, Seth McGinnis, Kenton McHenry, Nazila Merati, Matthew Miller, Subhashree (Shree) Mishra, Gretchen Mullendore, Fiona Murphy, Greg Nawrocki, Sawyer Newman, Tran Nguyen, Eric Nienhouse, Kevin Paul, Chuck Pavloski, Mohan Ramamurthy, Niall Robinson, Sarah Ruth, Russ Schumacher, Doug Schuster, Mena Shah, Shelley Stall, Greg Stossmeister, Don Stott, Gary Strand, P. Obin Sturm, David Tarboton, Masako Tominaga, Julian Turner, Kevin Tyle, John VanDecar, Jianwu Wang, Anne Wilson, Steve Worley, Fuqing Zhang

Remote Participants: Raleigh Martin, Ouida Meier

Workshop Steering Committee Members: Robert R Downs, Sophie Hou, Danie Kinkade, Matt Mayernik, Tran Nguyen, Doug Schuster, Greg Stossmeister, Fuqing Zhang

10. Appendix III - Workshop Agenda

The Workshop agenda, which includes links to plenary presentations and notes by theme and breakout sessions, can be accessed online at: https://www2.cisl.ucar.edu/events/workshops/geodarrs-workshop/2018/agenda

GeoDaRRS workshop agenda 1:00 PM, 8/7/2018 - 12:15 PM, 8/9/2018

NCAR Center Green Campus, CG1-1210-South-Auditorium, 3080 Center Green Drive, Boulder, CO 80301

- **8/7/2018, Tues PM 1:00-5:25 PM : What is the existing landscape and what gaps exist in that landscape for data producers and users? -Moderator, Doug Schuster, NCAR**
    - **1:00 - 1:30 Participant Registration.**
    - **1:30 - 1:40 Workshop introduction, Doug Schuster, NCAR**
    - **1:40 - 1:50 Review pre-workshop survey results, Matt Mayernik, NCAR**
    - **1:50 - 2:05 Building Community Informed and Driven Data Services at NCAR: Accomplishments and Roadmap, Sophie Hou, NCAR**
    - **2:05 - 2:20 BCO-DMO: Domain-Specific Data Management Services for the Marine Biogeochemistry and Ecology Communities, Danie Kinkade, BCO-DMO**
    - **2:20 - 2:35 Globus Platform Services for Data Publication, Greg Nawrocki, Globus**
    - **2:35 - 2:50 The Neotoma Paleoecology Database - Current Infrastructure, Ongoing Challenges, and Future Directions, Jessica Blois, UC Merced**
    - **2:50 - 3:05 FAIR-aligned Scientific Repositories: Essential Infrastructure for Open and FAIR Data, Shelley Stall, AGU**
    - **3:05 - 3:10 Summarize breakout group charges**
    - **3:10 - 3:30 Break**
    - **3:30 - 4:45 Breakout groups sessions**
        - **South-Auditorium, 2503, 2603 -remote participation available, 3131**
    - **4:55 - 5:25 Breakout group reports and preview tomorrow**
- **8/8/2018, Wed AM 8:00 - 12:15: What would you like to have and why? - Moderator, Tran Nguyen, UC Davis**
    - **8:00 - 8:30 Participants arrive**
    - **8:30 - 8:40 Introduction to the day**

    - **8:40 - 8:55** Pangeo's vision for scientific computing in the big-data era, Joseph Hamman, NCAR RAL
    - **8:55 - 9:10** HydroShare: A Platform for Collaborative Data and Model Sharing in Hydrology, David Tarboton, USU Logan
    - **9:10 - 9:25** A deep submergence vehicle user's perspective about data management services at NSF-supported facilities, Masako Tominaga, Woods Hole Oceanographic Institution
    - **9:25 - 9:40** The challenges and opportunities of being able to interrogate ensembles of numerical weather prediction models, Russ Schumacher, CSU Fort Collins
    - **9:40 - 9:45 Summarize breakout group charges**
    - **9:45 - 10:15 Break**
    - **10:15 - 11:30 Breakout group sessions**
        - South-Auditorium, 2503, 2603 -remote participation available, 3131
    - **11:45 - 12:15 Breakout group reports**
    - **12:15 - 1:30** Lunch
- **8/8/2018, Wed PM 1:30 - 5:00 PM:** What is realistic/doable with constraints? **-Moderator,** Robert Downs, CIESIN, Columbia University
    - **1:30 - 1:35 Intro to afternoon sessions**
    - **1:35 - 1:50** Funder perspective -NSF funding pathways for data related activities and research, Subhashree (Shree) Mishra, NSF
    - **1:50 - 2:05** Science perspective -Big Data, Fuqing Zhang, PSU State College
    - **2:05 - 2:20** Science perspective -What is realistic and doable for an atmospheric chemistry database? , Tran Nguyen UC Davis
    - **2:20 - 2:35** Data repository management in the environmental sciences in the UK, Sarah Callaghan, British Atmospheric Data Centre
    - **2:35 - 2:40 Summarize breakout group charges**
    - **2:40 - 3:10 Break**
    - **3:10 - 4:25 Breakout group sessions**
        - South-Auditorium, 2503, 2603 -remote participation available, 3131
    - **4:30 - 5:00 Breakout group reports and preview tomorrow**
- **8/9/2018, Thurs AM 8:30 - 12:15:** How do you move forward and build upon this discussion with your community? **-Moderator,** Danie Kinkade, BCO-DMO

- **8:30 - 9:00 Participants arrive -Light breakfast provided adjacent to CG1 South Auditorium**
- **9:00 - 9:30 Review and prioritize outputs from day 1 and 2**
- **9:30 - 9:45 Advancing the culture of data sharing at the U.S. Geological Survey through community engagement, Leslie Hsu, USGS**
- **9:45 - 10:00 New Ways to Deal with Data in the UK Met Office, Niall Robinson, UK Met Office Informatics Lab**
- **10:00 - 10:15 Building an Environmental System Science Community Data Archive, Deb Agarwal, Lawrence Berkeley National Lab**
- **10:15 - 10:30 What Do We Do Next?, Ken Bowman, Texas A&M**
- **10:30 - 10:45 Break**
- **10:45 - 11:45 Breakout group sessions**
    - **South-Auditorium, 2503, 2603 -remote participation available, 3131**
- **11:45 - 12:00 Breakout group reports**
- **12:00 - 12:15 Review follow on steps for workshop report and requested participant actions.**

11. Appendix IV - Breakout Session Discussion Questions by Workshop Theme

Breakout Session 1: What is the existing landscape and what gaps exist in that landscape for data producers and users?

1. Where do you currently archive the data that you produce?
    a. Why do you use this/these location(s) for archiving?
        i. Specific factors that motivate this choice (e.g. specific characteristics/capabilities of repository)?
    b. What other "go-to" location(s) do you have/know for archiving your data outputs?
    c. How did you learn about these archiving locations?
    d. What are your reflections on the outcomes of our survey?
2. How do find data to support your research?
3. How do you obtain data to support your research?
    a. Do you have a "go-to" location to obtain data to support your research needs?
4. What requirements do your funders and publishers have for open data access?
5. What guidance are you using that is currently provided by funders and publishers to support open data access?

Breakout Session 2: What would you like to have and why (resources are not an issue)?

1. From a data perspective, what needs to be done better to facilitate scientific discovery?
2. Do you value data management planning? Is there any value to you with this activity? What motivates you to do this?

Breakout Session 3: What is realistic/doable with constraints?

1. What are the largest challenges you encounter when sharing your data?
2. What are the concerns you have that discourages you from sharing data with others?
3. How should responsibilities be distributed amongst the various stakeholders?
    a. Who are the stakeholders?
4. Where should "data management" efforts be focused?
    a. Only future projects?
    b. Resurrect legacy data from past efforts (tied to publications)?
5. What is the relative value that could be derived from focusing on either effort (relative to future scientific discovery)?

Breakout Session 4: How do you move forward and build upon this discussion with your community?

1. Based on our workshop discussions:
    a. What is needed?
    b. What would be most helpful?
    c. What would you use?
2. What's the most realistic way to move forward and keep your community engaged and who should lead the conversation?
    a. E.g. Professional societies (AGU, AMS, other?), Funders, Publishers, Science community advocates
3. Would it be useful to have follow on workshops? How could they be done differently

12. Appendix V - Summary of Notes Organized by Workshop Theme

- ***Theme #1: What is the existing landscape and what gaps exist in that landscape for data producers and users?***
    - Existing Landscape:
        - There is a wide range of stakeholders, but each with specific missions and scopes (and limitations).
        - The current landscape also has many uncertainties, including what researchers should and can use.
        - However, a number of useful and successful existing domain services, which are developed by the community, are available.
    - Gaps:
        - Since there is no “one-size-fits-all” solution, understanding who can do what (perhaps a mapping of relevant roles and responsibilities?) and more consistent guidance could be helpful. In particular:
            - Researchers would like to have options when it comes to choosing data services and fulfilling data management activities, but the learning and selection process could be more beneficial if it is guided.
                - Additionally, the approach should be consistent among funders and publishers while being flexible enough for researchers in different situations to be able to meet the requirements separately.
            - Also, researchers would like to have clearer guidelines from both funders and publishers regarding what research outputs (including types and granularities) should be archived and for how long in order to meet the data management requirements.
                - Different data types (e.g. models vs observations) have different characteristics and needs, so they will need to be supported accordingly.
                - Especially in terms of sensitive data, who is responsible in deciding these issues?
    - General Concerns:
        - Sustainability of repositories.
            - This includes sustaining data storage availability/cost, so that these factors do not become constraints.
            - What is the balance between domain specific vs. general purpose repositories and how can the funders help in determining this?
        - What if a domain does not have existing data management practices; how to foster and facilitate support?
        - How can repositories provide services for the researchers in such a way that can help in making research more productive?
        - Motivations to participate in data management activities.
            - A mixture of “carrots and sticks” is needed.
                - This might include exploring how peer review process can be incorporated into data sharing process. Is this a training, willingness, funding, or other issue?
        - How can funding for data management activities be made available for both during and post projects?

            - Can researchers hand their products (e.g. software and data) off to a separate resource to curate their products once a research project is complete?
        - What are some services that could be appropriate to be deployed via the cloud?
        - How about legacy data? How should it be prioritized?

- ***Theme #2: What would you like to have and why?***
    - Guidance for what to save and where to save data both for meeting funders and publishers' requirements as well as for promoting scientific advancement for the long term.
    - Funding support for data management activities.
        - This includes:
            - Increasing funding cap per project to cover the costs of storage.
            - Improving the user experience of the repositories.
    - Provide data management resources and education for <u>all levels</u> of researchers on topics including:
        - Develop/implement/integrate education for researchers to learn about data management best practices.
        - Integrate data management planning and data professionals into research workflows starting from the beginning of the projects and with funding provided.
        - Provide mechanisms for:
            - Discovering what data management planning educational materials as available by discipline (e.g. best practices, standards, tools, etc.)
            - Sharing lessons learned among institutions/among disciplines.
            - Finding an appropriate repository for your data (repository directory or map?)
        - Compile and share use cases of what works and what did not work, including demonstrating how data management planning helps science.
    - Train data professionals as well as promote career paths and the value of the roles, especially in terms of contributing to research work.
    - Improve data management plan's structure both in terms of the processes for creation (e.g. data management plans should be interactive and include quality assessment.
        - Involve repositories in both processes.
    - Leverage new technical capabilities, such as cloud, so that "things can/should just work".
        - This includes:
            - Being able to combine data easily from various discipline ("convergent research").
    - Funders to work not just with their community but also with each other (both nationally and internationally) to be more consistent with their message/position/guidance and funding support.
        - This would also help in minimizing waste, improving interoperability of data shared, and providing positive reinforcement of funders' commitment to data management.
    - Funders to provide feedback mechanisms regarding data management policies/requirements/needs.
    - A resource to manage outputs/products (data, software, etc) once grants are complete in order to:
        - Store large amount of data (e.g. model outputs) with sustained funding.
        - Ensure data products are not lost because there is no support beyond the funded

cycle to take care of the outputs for the long term.
        - Facilitate commitments to continuity (both in terms of infrastructure and knowledge).
        - Hydrology (water modeling), Atmospheric Science.

- ***Theme #3: What is realistic/doable with constraints?***
    - Clarification of roles and responsibilities. This includes:
        - Clarifying who should be responsible and can help with preparing data to be archived and answering questions about the data after it has been shared.
            - The preparation and ongoing maintenance work could take significant amount of time.
        - Linking researchers with resources, including human expertise, for data management, analysis, workflows, tools, etc.
        - Having consistent institutional policies and guidance and enabling tracking/accountability at institution level.
        - Providing guidance to facilitate discussion and understanding/agreement in advance, especially during proposal development phase.
        - Enabling complementary funding models (e.g. funder directly pays for repository services, allow an line item to be included in the proposal, or?)
    - Licensing/intellectual property issue is crucial to deal with social issues related to sharing datasets.
        - Features need to be in place to deal with intellectual property issues, including credential verification and access control guidance?
    - Technical barriers of sharing large datasets - what needs to be saved?
        - Infrastructure limitations, along with lack of human resources, can also be a hindrance for fulfilling data management activities.
    - Data management training, especially for scientists, should be provided earlier on, but interest/time might be limited for scientists. Ideas for providing data management training/education include:
        - Identifying groups that could benefit from the training.
        - Providing internship opportunities supported by NSF.
        - Integrating with standard undergrad and grad school scientific curriculum.
        - Further defining the data professional roles, titles, and the education/training requirements to become such professionals.
    - Reproducibility is crucial, but there are many questions.
        - What is the definition of reproducibility and granularity?
        - What are the goals for reproducibility?
    - Stakeholders:
        - Funder, researcher (provider), researcher's institution, publisher, repository, data user/reuser, public (taxpayer), professional societies, storage provider, repository platform provider, commercial sector, data professionals (data curators/data manager/librarian/data analyst/data modeler/data miner).
    - It is really critical to understand the funding question; particularly:
        - Who pays for what in order to meet the requirements for the short term and the long term?
    - Effort should be split to provide services for both future and legacy projects, but emphasis can be placed more on future projects.
        - Perhaps 80% future and 20% legacy or as required by evolving science needs.
        - Referencing legacy data in current publications can be a motivator for

curating/archiving legacy datasets
    - NSF needs to show that it cares about the data management planning process and establish trust.
        - For example, NSF might be able to demonstrate this by:
            - Facilitating the sharing of resources and services, such as open source software, quality and reusability information, cost models for archiving datasets, and skill sets and expertise.
            - Holding data management workshops
            - Providing consistent feedback to PIs on the data management plans.
            - Having better guidance for new PIs.
            - Providing well defined data management plan review guidelines for proposal reviewers, and elevate the importance of complete data management plans in the review process.

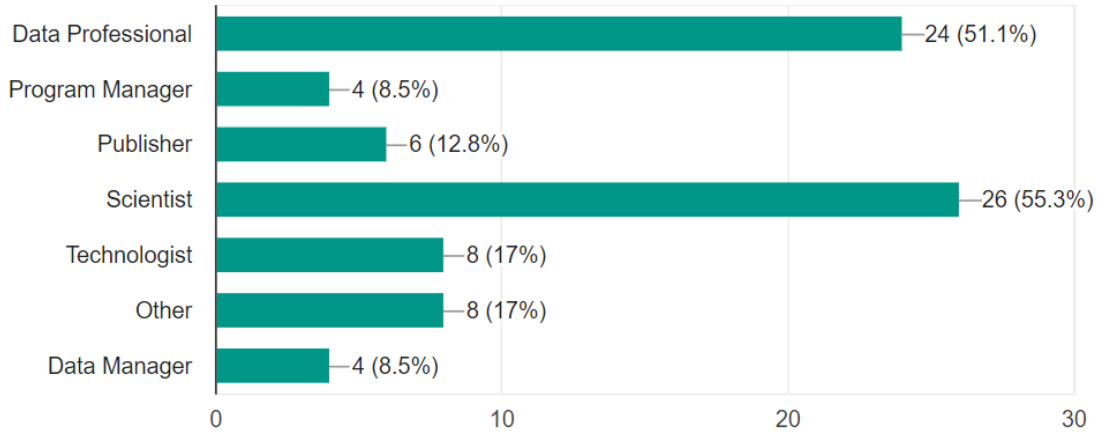
- ***Theme #4: How do you move forward and move upon this discussion with your community?***
    - Draft the report and publications to keep participants engaged.
    - Request responses from NSF about our recommendations.
    - Outreach and engagement activities:
        - Workshop of leaders in modeling field and related disciplines.
        - Hackathons (e.g. scientists and data managers).
        - Work with the professional societies to help in sharing knowledge with colleagues, especially in terms of the value of the data professionals and data management.
        - Coordinate cross pollination across CI and early career committees in professional societies.
    - Leverage PIs to continue discussing with their communities.
        - Continued discussions with scientists is needed.
        - More discussions among scientists and repository managers are also needed, especially in terms of sharing best practices and guidance.
        - Identify champions in disciplines to help in sustaining the discussions and exchanging feedback.
    - Need to push for the repositories to be part of the DM planning during the proposal process, especially from funders.
    - Having NSF review DMPs consistently and bring in repositories in the discussions earlier.
    - Understand how NSF can evolve its own culture (ex: collaborate among the directorates, improve funding cycles, train the proposal/DMP reviewers and clarity, etc.)
    - Need to be aware of the full spectrum of data (There's a lot of data types that can fall through the cracks between observations and models).
    - Research Coordination Network grants proposals to focus on specific issues.
    - Publishers are starting down a good path with FAIR initiative, but they have to be flexible when crafting policies
    - Leverage social media and GitHub?
        - For social media could it be used to bring awareness to data management and share reviews/ratings of data services/
    - Review emerging technologies (e.g. cloud services, Jupyter Notebook, etc.).

13. Appendix VI - Pre-Workshop Survey Results

Note: Free-text “Other” responses are summarized for any questions where “Other” was the most common response.

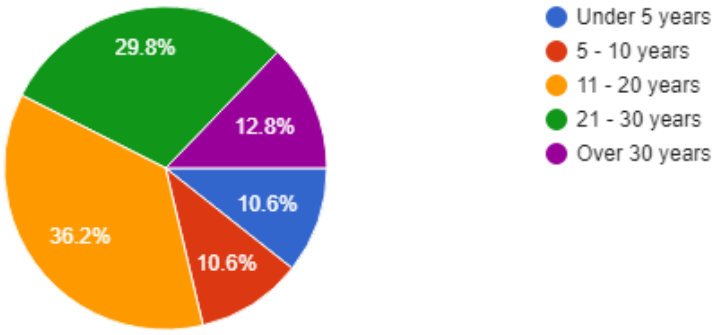
Your Primary Role(s)/Responsibilities (please check all that apply)

47 responses



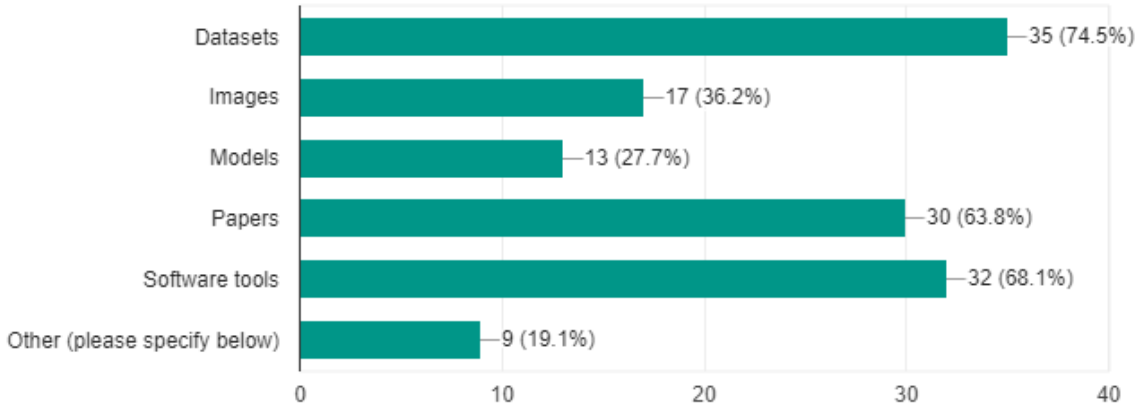
Your Years of Experience

47 responses



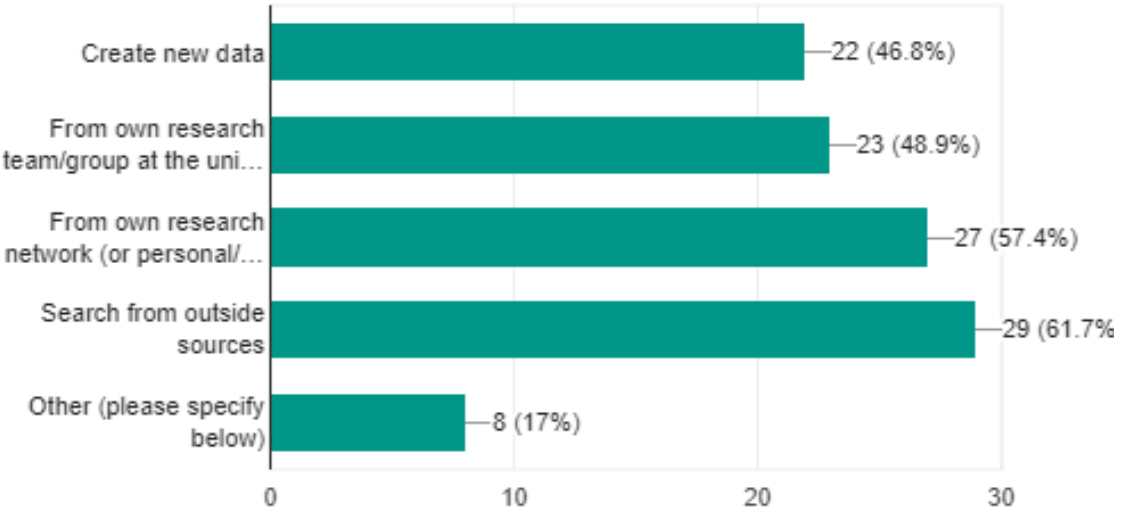
What types of outputs do you produce from your research? (please check all that apply)

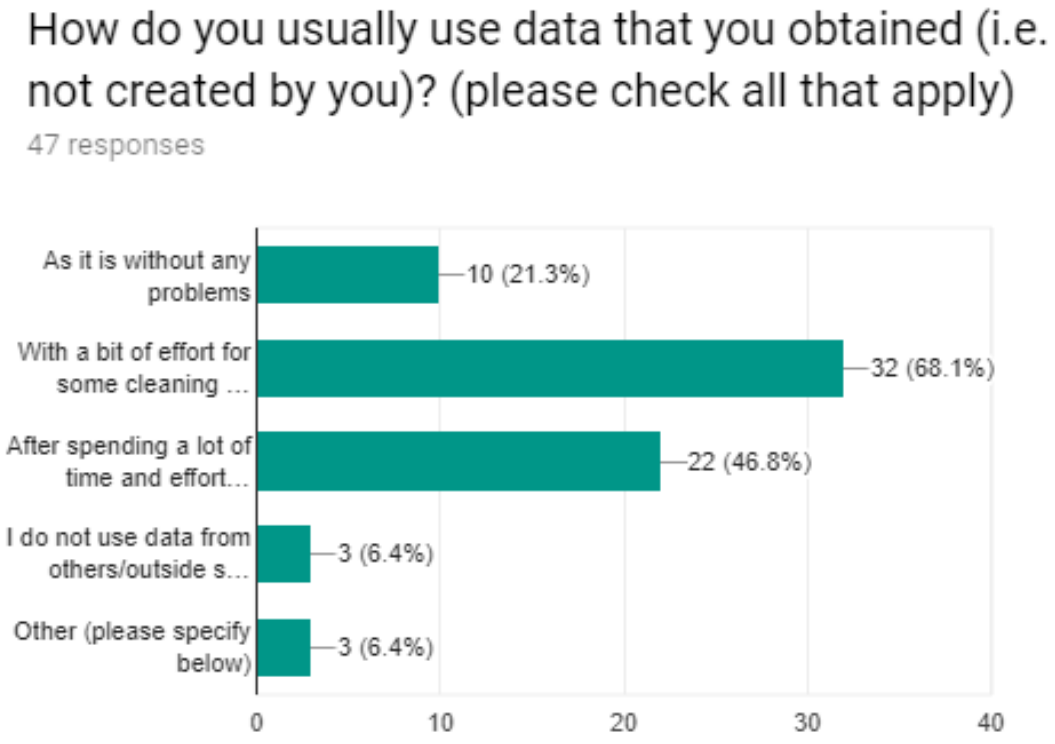
47 responses

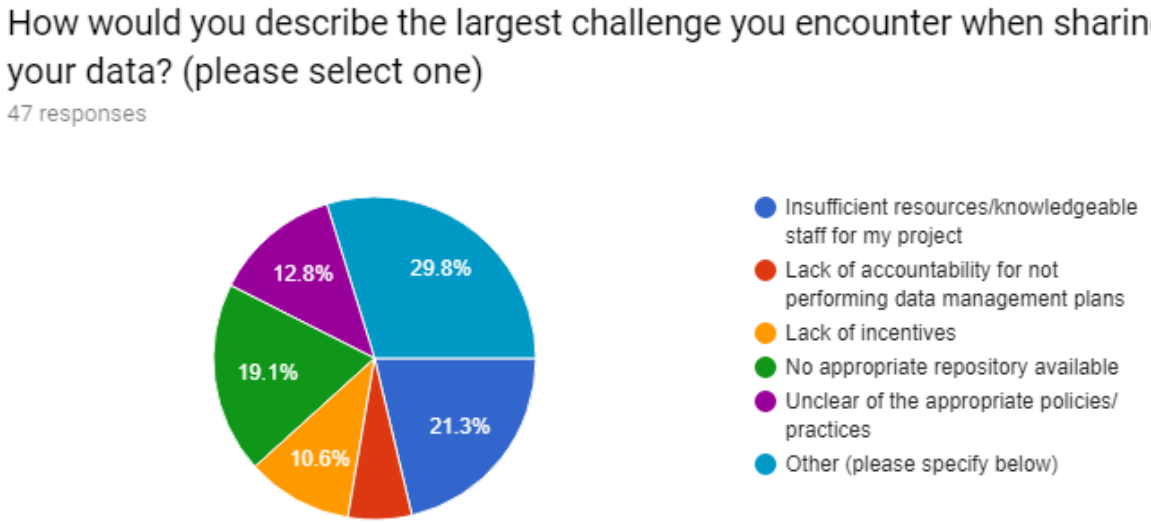


How do you usually find and get the data for your research? (please check all that apply)

47 responses



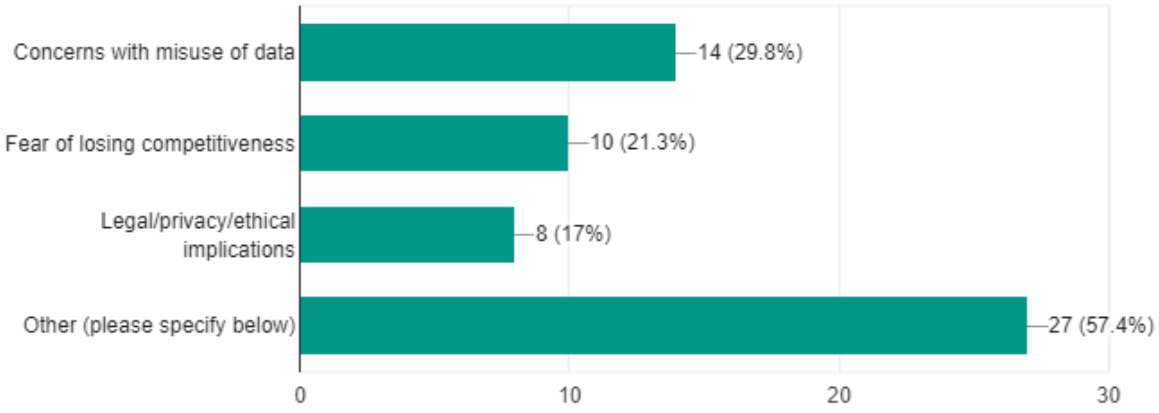




Summary of "Other" responses. Free text responses were related to the following issues: Not applicable (4), Data volumes and complexity (4), Model data and data processing software (4), Metadata and data organization (3), Lack of time to prepare and/or document data (2), No challenges (1), Overwhelming number of data requests (1), Data access services (1)

What are the concerns you have that discourages you from sharing data with others? (please check all that apply)

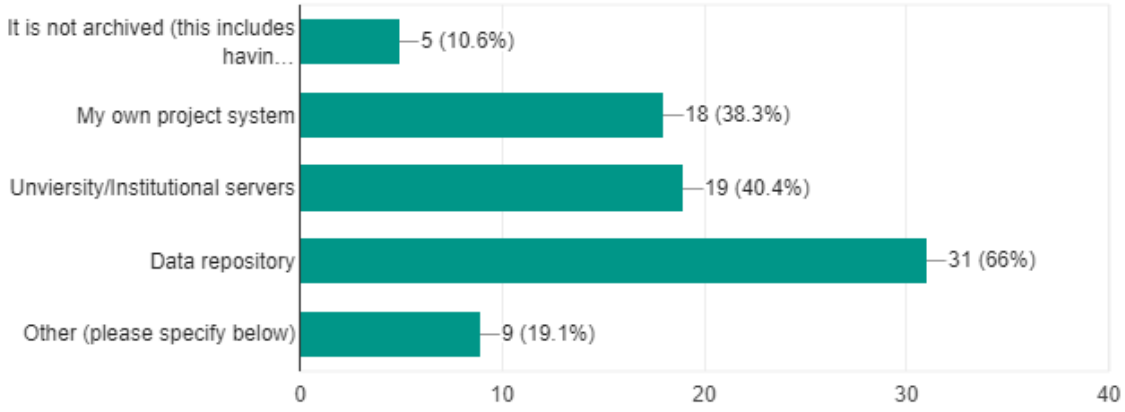
47 responses

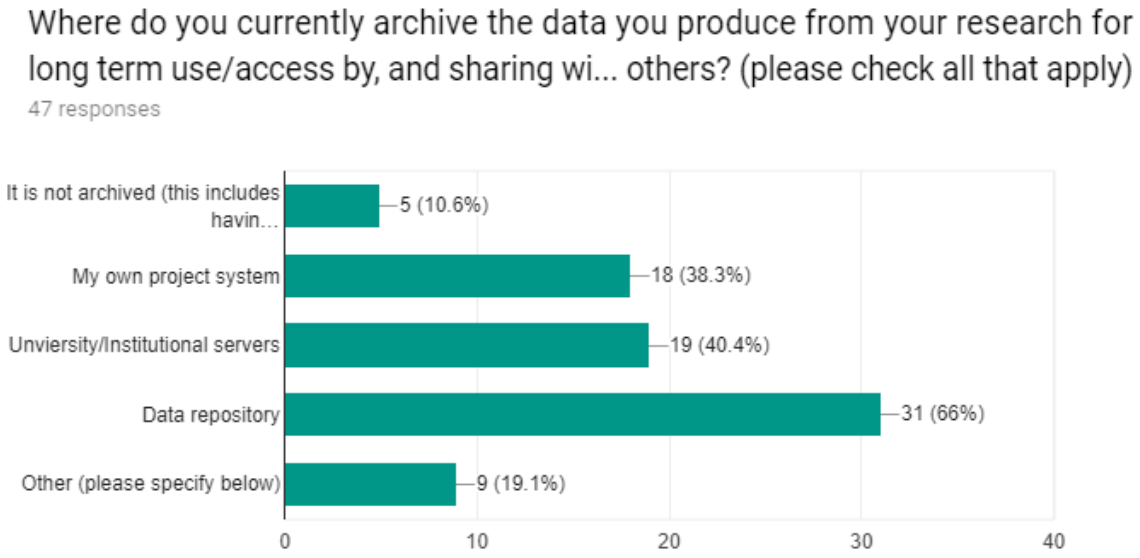


Summary of "Other" responses. Free text responses were related to the following issues: none (6), not applicable (6), Time and/or effort required (6), Lack of institutional support (4), Misuse or misunderstanding (2), Data accuracy and/or quality (2), Loss of competitiveness advantage (2), Unclear on what to share (1)

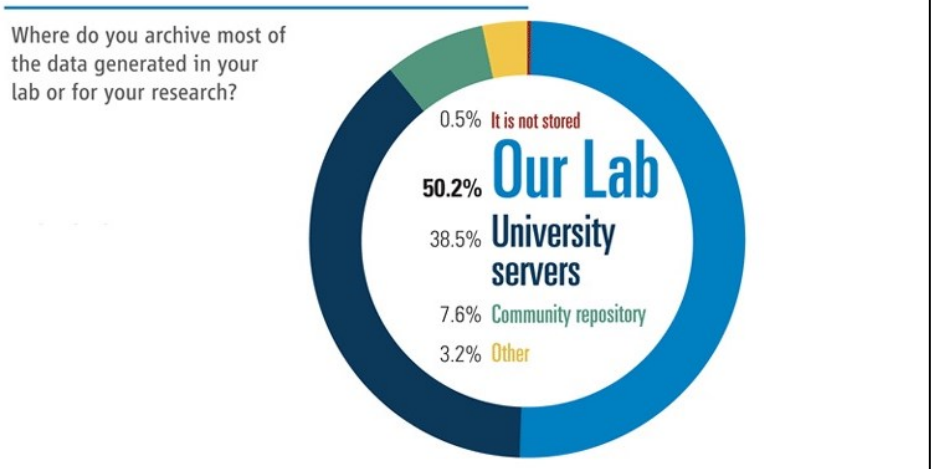
Where do you currently archive the data you produce from your research for long term use/access by, and sharing wi... others? (please check all that apply)

47 responses



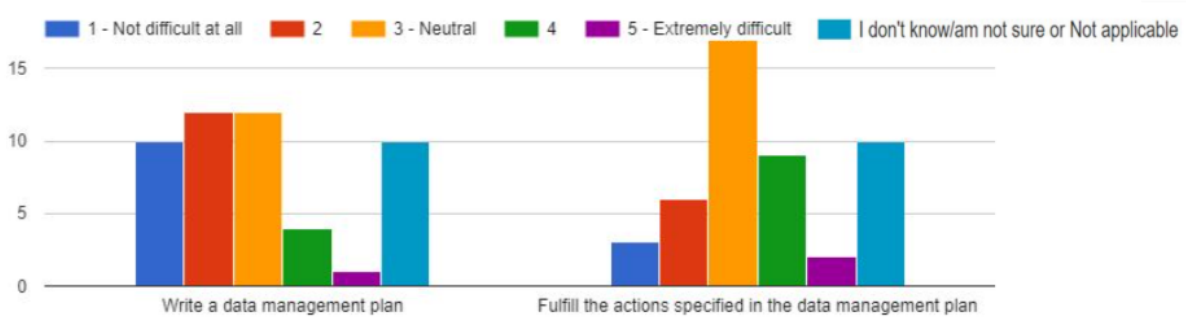


The above question was taken from a survey of 1,700 scientists conducted in 2011 by *Science* magazine. The chart below is reproduced from the Science survey results as a point of comparison.



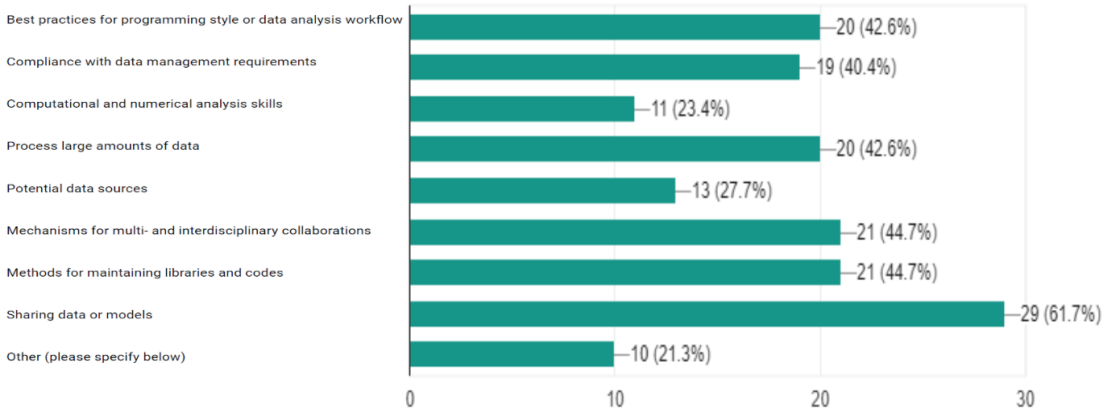
Science Staff. (2011). Challenges and Opportunities. *Science,* 331(6018): 692-693. http://dx.doi.org/10.1126/science.331.6018.692

On a scale of 1 to 5, how difficult is it for you to: (please select only one value per row)

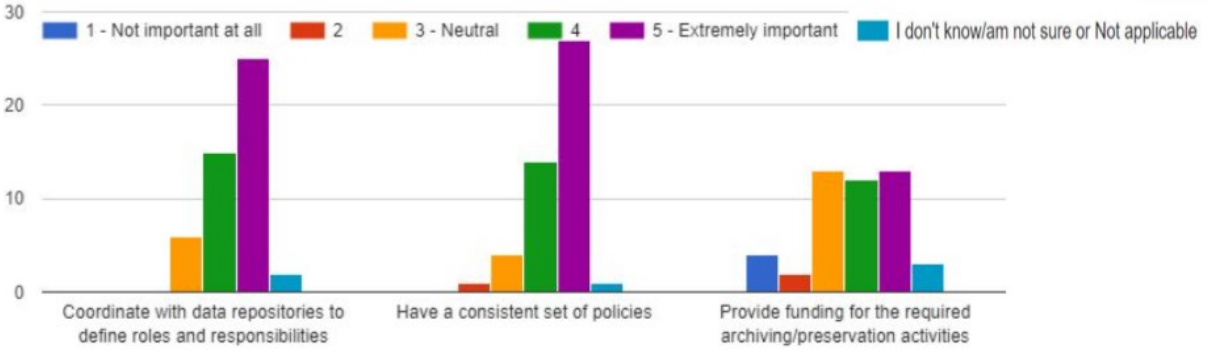


Which research habits and digital skills would you like to receive support/training in? (please check all that apply)

47 responses



How important is it that the PUBLISHERS support data repositories in the following areas: (please select only value per row)



On a scale of 1 to 5, how important is it that the FUNDERS support getting data into repositories by: (please select only value per row)

14. Appendix VII - Summary Table of Workshop Recommendations by Topic

Topic	Challenge	Recommendation	Desired Outcome
Long-term, Scalable Data Curation	Researchers are required to document, archive and cite data by funders and publishers to support open access requirements. Many researchers do not know how or if resources, such as repositories, are available to meet these requirements. Projects that generate large data volumes, such as model outputs, face additional resourcing issues. At present, individuals dealing with these issues develop temporary, ad hoc solutions, including putting data on cloud infrastructures, local research group servers, and university-operated servers. In many cases these ad hoc solutions aren't sustainable after short term grants end.	Long-term support for the data curation needs of the geoscience community is critical for providing truly open access to data. Several different approaches to providing sustainable open access to data are possible: 1. Augment existing geoscience data repositories to scale up their capacity. 2. Identify non-specialized data repositories that fulfill open access objectives. 3. Develop a data repository liaison service. 4. Create new data repository services.	Researchers can hand off data products to a separate resource at the end of a project. Researchers meet open access requirements. It is easier to build upon the research and data products of others. This will foster new scientific discovery. Repositories have sufficient resources to effectively curate and provide access capabilities to large volume dataset collections.
Education & Training	Many researchers are not aware of best practices for data management, and do not include these best practices as part of their research workflows.	There is a need for programs to better support data management education within scientific, computing, information, and data disciplines. Data management training and resources for researchers need to be improved, and better publicized.	Data management best practices are integrated into scientific research workflows. The burden of meeting open access requirements is eased.

Topic	Challenge	Recommendation	Desired Outcome
Data Management Plans (DMPs)	There is a perception in the scientific community that data management plans are an “afterthought”. There is a perception in the scientific community that there is little accountability for following through on a data management plan. There are numerous templates available for data management plans, but insufficient guidance within the geoscience community on which templates to use. Repositories have difficulty in resourcing for incoming products when not included at the project planning phase.	Grant proposal reviewers should 1) review data management planning according to multiple explicit criteria including sufficiency, resourcing, and execution plans, and 2) scrutinize this section as critically as all other sections of a proposal. An efficient mechanism for grantees to update and comment on their DMPs during the annual reports would help improve the accountability for the DMPs. Data repositories need to be brought into the DMP conversations in the initial stages of the project planning process. The geoscience community should foster a DMP tool ecosystem.	Data management plans have consistent expectations, and are consistently structured and reviewed. Researchers follow through on executing their data management plans without undue burden. Researchers are guided to the repository that is best suited to curate their data products. Repositories resource their facilities appropriately.
Funder & Publisher Policies	Inconsistent data management policies across funders and publishers create a complex landscape for researchers to navigate. This can lead to challenges in publishing research findings, and fulfilling funder data management requirements. It may be impractical for projects that generate large data volumes, such as model outputs, to retain the full data record to support open access requirements.	All stakeholders should be clear on the core drivers and principles that motivate their data policies. Coordination among all stakeholders is necessary to bring consistency to the data policy landscape. Specific scientific research communities need to discuss and formalize data retention guidelines.	Researchers have well-defined pathways to meet both funder and publisher data management requirements. Funder and publisher data management requirements are reasonable in terms of scope and effort asked of researchers and repositories. Data management related tasks are easier for the research community to achieve.

Topic	Challenge	Recommendation	Desired Outcome
Strategic Partnerships	Many federal agencies, universities, and research institutions face similar data management challenges, yet build disconnected, insufficient and/or independent solutions. This can lead to duplication of effort and spending. Purchasing of resources such as compute, storage, and cloud-based solutions can be prohibitively expensive for individual projects or institutions.	Strategic partnerships across federal agencies could reduce costs through shared data storage and curation services. Strategies need to be developed at the agency level to employ cloud computing and storage.	A broader set of services, including public cloud supported data storage, access, and scalable analysis, are provided across repositories from different disciplines. This could facilitate more efficient interdisciplinary research and discovery, and increased access by a more diverse user community, including the commercial sector. Duplication of effort across agencies is minimized, leading to cost savings.
Legacy Data	Legacy data created/collected through past projects exist throughout the research community. In many cases, these data do not adhere to the format and metadata requirements needed to support long-term curation. When the grants used to create/collect these legacy data expire, researchers have few options for data “cleanup”, storage, or archiving.	Researchers need clear paths to support curation and rescue of data collected via past projects.	Researchers have the capability to resource “data rescue” efforts in addition to current projects. Legacy “dark data” are brought up to curation level standards and added to repositories. These data could assist in answering new, or lingering research questions.

Topic	Challenge	Recommendation	Desired Outcome
Tools & Services	Scientists push the boundaries of existing tools, often customizing or extending tools in unique ways. “Off the shelf” software is thus not always sufficient as a data analysis solution. Data storage costs continue to be an ongoing challenge. Despite the decrease in storage costs on a per-unit basis, total storage costs continue to increase as the amount of data generated through computational models and new observational instruments and sensors also increases.	All stakeholders should recognize the importance of open source software communities, and contribute to these efforts where possible. Data repositories should investigate whether cost efficiencies can be gained by sharing data storage infrastructure.	Community supported open source software solutions will better meet the data analysis needs of a broader set of research communities. Researchers that contribute to open source software efforts are more likely to transition to these and other related new technologies since they have invested in their development. Repositories can focus more resources on data curation related activities versus storage hardware. Data archival, access, and analysis tool development across repositories could be streamlined by using a common storage infrastructure. This would result in a more consistent end user experience across different repositories as well.