Research Data Management Implementations: Impact on Science

Preservation, Reuse, and Reproducibility

NSF-Sponsored Workshop Report

Final report for the National Science Foundation-sponsored workshop, held at the Westin Arlington Gateway in Arlington, Virginia on September 14 and 15, 2017.

DISCLAIMER

The workshop described in this report was supported by the National Science Foundation under the NSF Award ID 1661523. Any opinions, findings, or conclusions expressed in this report are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Table of Contents

Executive Summary

Since the codification of the scientific method in the 17th century, the proper management of data has been at the heart of research and discovery. Scientific results are communicated in part through the publication of data and results, collaborative and progressive advances are made through the sharing of data between researchers, and verifying the legitimacy of a new finding requires the comparison of new data with old.

Though today's scientific results are measured in terabytes instead of notebook pages, research data management remains central to the practice of modern science. If anything, data management is now even more critical, as nearly every field is swept up in the "data deluge" of rapidly increasing data volume and complexity. Yet as the scientific community slowly accepts these truths, many challenges remain in the proper implementation of tools and solutions that support best practices.

This workshop was the latest in a series bringing together a cross-section of 90 research professionals to discuss the progress and persistent cultural, technical, and funding obstacles of today's data management implementations. Researchers and representatives from university IT and computing centers, libraries, funding agencies, journal publishers, and industry met for two days of talks, panels, and small-group discussions on these topics. This report summarizes the output of those activities, as well as the findings of position papers submitted for the workshop.

Much of the conversation was underscored by optimism about the growing number of scientists realizing the importance of research data management, driven by factors such as new funding agency policies and increased awareness of data resources. However, many participants also pointed to a difficult gap between research enthusiasm for proper data management and the actual execution of those practices, due to a lack of funding, institutional guidance, uniform standards, and incentives.

Findings

- Researcher usage of Research Data Management (RDM) is increasing, driven by greater awareness of its importance among younger scientists, funding agency requirements for a data management plan in grant proposals, widely available and user-friendly tools, and the overall shift toward data-driven research in most research fields.

- However, as demand for these services increases, critical gaps remain in funding, staffing of trained research data professionals, and training of researchers in data management skills.

- The rapid expansion of the research data management tool ecosystem continues, with better solutions for data storage, sharing, and discovery emerging on a regular basis.

Going forward, careful scientific assessments are needed to identify those that work best for users, so that those tools can be more widely promoted and utilized.

- Reproducibility is a key tenet of the scientific method, and research data management is crucial for reproducibility in modern science. Publication and sharing of data and other computational research products (e.g. software, workflows) enable the verification of research results.

- Libraries increasingly play an important role in research data management on campus, providing technical training, curation services, and assistance with assembling data management plans for grant applications. Whenever possible, campuses should coordinate RDM services and strategy between IT services, libraries, computing centers, and research faculty, as it avoids redundant spending and effort and utilizes expertise across the institution.

- Data management plans are quickly becoming mandatory for grant proposals to federal funding agencies. NSF created its data management plan in 2011, the NIH is currently reviewing policies to make these plans mandatory, and NOAA and NIST both require plans of some form for their intramural research. These requirements have driven up participation in research data management; however, the format of these data management plans and the standards for their peer review remain in flux.

- Businesses such as Elsevier, Amazon, and Microsoft are focusing their research data management efforts on creating tools for data discovery and sharing, creating new platforms, collaborative data commons, and elastic edge computing technologies.

- There is an intense need for data standards to ensure portability, interoperability, and easier data discovery. Even within single disciplines, consensus rarely exists on standards for data and metadata formats, and there is no agreed-upon authority to set those standards.

> **"We are all astronauts now. Everybody here has a smartphone that exceeds all of the technology on the Voyager satellite. There's a huge amount of data and tremendous challenges and questions to be addressed." - Amy Walton, Program Director for the Office of Advanced Cyberinfrastructure, National Science Foundation**

Recommendations

- At the campus level, more administrations need to commit to supporting RDM resources and determine a sustainable strategy for long-term data storage, curation, and sharing.

These conversations should involve all campus stakeholders (IT, library, researchers, computing centers) and determine governance, policies, and budgets for campus data management.

- In addition to research data management implementation workshops such as this one, create more RDM workshops at discipline-specific meetings, run by people with experience in data management but aimed at researchers who are new to the topic.

- Incorporate research data management training into all levels of curricula to reach new researchers and educate them that data management is an essential aspect of modern science.

- To appeal to senior faculty, find campus “champions” who can speak to the benefits of data management and help steer researchers to the proper resources at the institution. Also consider different formats of training, including webinars or online modules, to draw in researchers less likely to attend an in-person workshop.

- Create new incentives for data management at multiple levels. At academic institutions, consider dataset publication and sharing as part of the tenure and promotion process. National-level collaborations or agencies can encourage sharing by requiring it if a researcher wants access to others’ data.

- Developers of research data management tools should focus on automation, “hands-free” functions that handle data during the natural workflow of science, instead of adding effort and time. These features increase adoption and compliance with data management standards, while also reducing cost and training demands.

- Hire additional data management professionals with expertise in science, IT, library curation -- if you can find them. Domain-specific curators are also important, but schools lacking resources to hire experts in each domain can form partnerships with other institutions to share expertise.

Introduction

Workshop Motivation

Since the last NSF-funded research data management implementation workshop (RDMI) in 2013, there have been many funded projects, research communities and organizations discussing and developing tools and services related to RDMI. This RDMI workshop was designed to focus on implementations of solutions for managing the storage of research data and encourage open discussion of available strategies for data management in specific fields based on actual case studies. Participants from academia, industry, and government gathered to reflect on what has been accomplished to date related to data management, sharing achievements and opportunities for further contributions moving forward. While there are research groups and scientific communities that have made investments in RDMI and greatly benefitted from their use, many researchers still face a number of challenges, such as what constitutes the appropriate RDMI for their research, what technologies are required, and how to secure the required skills and support. These challenges are also driven by the increasing size, complexity, and velocity of data from research instruments, observations, experiments, and simulations that need to be managed and preserved.

Workshop Objectives and Participation

To build upon the foundation of the 2013 RDMI workshop, the University of Chicago Research Computing Center, in partnership with the Coalition for Academic Scientific Computation (CASC), organized a second Research Data Management Implementation workshop. This workshop focused on: (a) the impact that RDMI has on science, (b) shared strategies for developing RDMI in partnership with research communities, and (c) increased access to RDMI for the broader research community.

The RDMI workshop brought together 90 people from research computing communities, library communities, funding agencies, and leading experts in data management. The two-day workshop consisted of a keynote address, panels discussing various topics, and small breakout sessions. To promote discussions from different perspectives and expertise, specific table assignments were given to each attendee.

Position and Experience Papers

A total of 13 position and experience papers from 36 authors were submitted for the 2017 workshop. These 13 positions papers were supplemented by the 30 position and experience papers submitted for the 2013 RDMI, enabling readers to compare and contrast the evolution of RDMI over four years and perspectives from various stakeholders and communities. At the event, planning committee member Ruth Marinshaw, chief technology officer for research computing at Stanford University, summarized the contents of these submissions.

The largest group of submissions were experience papers, describing research data management policies and tools at different institutions. Kia Huffman, Tom Hoogendyk and Elizabeth Brainerd of Brown University wrote about their experience creating a data portal for motion analysis research which uses a 3D imaging technology that generates hundreds of files with each experiment, large file sizes, and complex metadata. Finding that the storage and repository tools provided by Brown were inadequate for their data, the group built the X-Ray Motion Analysis Portal, or XMAPortal, for data and metadata management and sharing, both within campus and to outside institutions and communities.

A similar challenge was covered in a paper by Adam Sierakowski of Johns Hopkins University, who described Craedl, the Collaborative Research Administration Environment and Data Library. This framework manages data for the Hopkins Extreme Materials Institute (HEMI), which consists of hundreds of researchers spanning multiple institutions and projects. Uniquely, Craedl incorporates research administration data management as well, so that researchers can track grants, projects, and publications in addition to experimental data.

Other tools described in workshop submissions included the GEMS (Genetic, Environmental, Management, and Socioeconomic data) platform from the University of Minnesota, created to support the International Agroinformatics Alliance with a focus on combining several different dataset types. The San Diego Supercomputer Center at the University of California-San Diego described the SeedMe platform, a more general research data sharing infrastructure that serves over 600 users and is currently undergoing an upgrade to its second iteration.

Another category of papers presented campus-wide solutions to deal with research data at scale. Carol Song, Preston Smith, Alex Younts wrote about how they have addressed the rapid expansion of data and computing needs at Purdue University, including a Community Cluster Program, a Research Data Depot, and the Research Environment for Encumbered Data. Hakizumwami Birali Runesha, director of the University of Chicago Research Computing Center, wrote about DaLI, the Data Lifecycle Instrument, a replicable platform for data management from acquisition through publication.

Beyond papers focused on specific technical data management solutions, some participants submitted experience papers on developing data storage strategies. Rajendra Bose and Bruno Scap discussed developing a data plan for a brand new center, the Zuckerman Mind Brain Behavior Institute at Columbia University. After engaging researchers at the center to understand their workflow, they decided to focus initially on “active storage” for day-to-day research tasks, with a longer-term focus on deep archiving.

Authors David Fearon and Reid Boehm from the Johns Hopkins University Libraries, described plans for updating their data infrastructure to address larger data files and increased researcher demand for data sharing and collaboration. Jamie Wittenberg of Indiana University discussed similar efforts to build a research data service in the context of a state university system with multiple campuses. Andrew Johnson of the University of Colorado Boulder, presented an

outcomes-based approach to research data management in contrast to a “one size fits all” approach.

A unique perspective was provided in a paper from Anita de Waard of the science publisher Elsevier, which describes the company’s Mendeley data management suite. In addition to information about this platform, the paper also contains valuable results of a 1200-person survey on data sharing practices, which found a significant gap between the desire of researchers to share data and the actual practice of data sharing. Respondents cited barriers such as lack of training, privacy concerns, ethics, and intellectual property rights.

A final subset of papers were structured around more general pressing challenges in research data management and potential solutions. Rachana Ananthakrishnan, Kyle Chard and Ian Foster of the University of Chicago and Argonne National Laboratory described two common “design patterns” pervasive across science — the research data portal and research facility automation — and how they can each be enhanced by elements of the Globus data management platform. University of Michigan’s James Myers asked “Where’s my Universal Data Browser?,” a call to action that described the aspirational vision of researchers finding and accessing any data relevant to their research, no matter where it was published.

Keynote: Research Data Management Implementations: Towards the Reproducibility of Science

The conference’s keynote speaker, Victoria Stodden of the University of Illinois at Urbana-Champaign, framed the state of research data management around a topic of much current discussion: scientific reproducibility. Numerous recent reports on the “reproducibility crisis” in science[1] have raised concerns about the reliability and transparency of scientific findings. In her talk, Stodden traced the origins of reproducibility in science back to the 1600’s, and discussed how today’s data-intensive research offers both new challenges and new solutions for reproducibility.

Independent verification of scientific claims has been a core scientific norm since at least the 1660s, when codified by Robert Boyle and the *Transactions of the Royal Society*. Nearly four centuries later, the same standards apply — at least theoretically — to empirical replication of new research results. By following the same methods, a second research group should be able to recreate the results of the primary authors, or at least a new set of data that supports the original conclusion.

However, the modern complexity of science makes this replication process more difficult than ever. Extremely large and high-dimensional datasets, sometimes produced through thousands

[1] http://www.nature.com/news/reproducibility-1.17552

or millions of hours of experiments, sensor activity, or computational simulation, are prohibitively expensive and impractical to recreate or share for re-analysis by an outside party. Traditional methods of communicating scientific work through static journal articles are ineffective for complex, computational research, with its reliance upon software and advanced statistical analysis.

Yet in other ways, today's technology and attitudes towards data make reproducibility easier. Data management tools, whether designed specifically for research or for more general uses, have lowered some barriers for data archiving and publication. Culturally, the "open data" ethos promotes the sharing of data among researchers for collaboration, replication, and additive usage, in contrast to established biases for data hoarding and secrecy.

Stodden argued that the increasing computational complexity of science can converge with the drive towards increased transparency for reproducibility, if the proper infrastructure and standards are developed. Reimagining the scientific article to include not just methods and results but also data, code, and workflows will both expand the scholarly record and enable others to verify and expand upon research claims. Such a system would promote good behavior, disincentivize poor methodology and statistical analysis, make it easier to compare computational pipelines and models, and unlock new discoveries.

> **"When you are running these much more computationally intensive projects, you're going to require a greater level of transparency just to keep track of your own research. The computational infrastructure necessary to run these bigger experiments is going to bring a level of transparency and organization naturally." - Victoria Stodden**

In a recent Science paper[2], Stodden and co-authors proposed several principles for computational science that could help fulfill this vision. These recommendations range from sharing data, code, and workflows in open repositories to proper citation and documentation of digital artifacts (such as software) to the usage of reproducibility checks and open licensing. Overall, these principles should support scientific norms of reproducibility and reuse, encourage best practices in research and discovery, and take a holistic, integrated approach to creating cyberinfrastructure that ensures interoperability.

Audience questions focused on how this vision would apply to data that is not easily shareable, such as data collected by private industry. Stodden predicted that the social shift towards open data now underway in academia will also spread to industry research, fueled by the need for scientific verification and the incentives of accelerated discovery with open, computational methods.

[2] http://science.sciencemag.org/content/354/6317/1240

Another audience member asked where these new standards should originate, given that different research communities are in different stages of acceptance for computational science and data management. Stodden[3] described it as a "collective action" problem with multiple stakeholders, including journals, funding agencies, and academic institutions. Cross-disciplinary discussions (such as the current workshop) were recommended to bring about convergence around similar principles.

Panel 1: Campus Support & Challenges with RDMI

The challenges of research data management touch many corners of the academic campus; not just the researchers themselves, but also IT services, librarians, and administrators. The workshop's first panel featured representatives from each of these roles sharing their current data management practices and struggles.

Multiple panelists observed increased demand for research data management services on their campus. Preston Smith, director of research services and support at Purdue University, presented graphs and statistics showing an exponential growth in data archiving and transfer needs and high-performance computing usage. Factors driving this trend include interest in reproducibility and data discoverability, rapidly accelerating data collection due to new equipment and methodologies, a need for continuity of knowledge as student and postdoctoral researchers leave the laboratory, and the desire for secure, long-term preservation of data.

Cultural shifts in science also drive increased interest in research data management. Juan De Pablo, Liew Family Professor of Molecular Engineering at the University of Chicago, observed that it is difficult to find a scientist under the age of 25 who doesn't believe computation is an essential part of research. New requirements from funding agencies for making data public and available have also created broader acceptance among researchers. The changing attitudes are reflected in increased demand for data management training, a service often provided by academic libraries and research computing centers.

But all panelists agreed that still more incentives, education, and innovation are needed to broaden participation in research data management. De Pablo urged the development of tools and infrastructure that requires little-to-no effort and time on the part of the researcher, making data storage a trivial part of the research workflow. At the University of Chicago, de Pablo's group collaborated with the Research Computing Center to develop the DMREF Datahub[4], a place for publishing both datasets as well as analysis and visualization tools, making validation simple. At Purdue, Smith's unit developed the Data Workbench[5], an interactive, in-browser

[3] You can access slides from Stodden's talk here: http://web.stanford.edu/~vcs/talks/RDMI-Sept14-2017-STODDEN.pdf

[4] https://datahub.rcc.uchicago.edu

[5] https://www.rcac.purdue.edu/compute/workbench

compute environment for data analysis and simulation that helps researchers ramp up to using high performance computing.

> **"We need to remind everyone of all the things we're *not* doing by being obsolete in managing data." - Juan de Pablo**

The panel also provided examples of organizational changes that could support and encourage data management. Rachel Vincent-Finley, Associate Dean for Academic Affairs at Southern University and A&M College, described LONI, the Lousiana Optical Network Infrastructure [6], which creates redundancy for data and computation at universities and government agencies in the state, protecting knowledge and research in case of natural disasters. Christie Ann Wiley, Engineering Librarian at the University of Illinois at Urbana Champaign, spoke about the library's Research Data Service[7], which provides expertise, tools, and infrastructure for the management and stewardship of research data. The library also helps researchers create data management plans and connects them with data publication resources such as the Illinois Data Bank[8].

Beyond these initiatives, the panelists agreed that more infrastructure is needed to continue shifting academic culture towards responsible data management. De Pablo argued that data storage should be viewed by schools as a utility, akin to air conditioning and power, that researchers are not expected to pay for. Audience questions drove discussion of who should bear the cost of building and maintaining that infrastructure: administration, researchers who use large amounts of storage or computing, or even publishers whose journals will soon require sharing of data.

Another discussion arose over whether, given limited resources, *all* data should be stored, and how long data archives need to be maintained. Smith and de Pablo argued that more storage is better, if not cost-free, pointing to current gaps in the research record such as data from unsuccessful experiments and graduate student dissertations. Panelists also raised concerns about how to clean or anonymize sensitive data so that it could be shared without compromising privacy.

Breakout Session 1: Today's Challenges in RDMI

In addition to panels and talks, the workshop included three breakout sessions where attendees formed into small groups and discussed pre-chosen questions about the current state and future of research data management implementations. Representatives from the groups then presented their responses to the full audience. For this report, we will provide the most common

[6] https://loni.org
[7] https://www.library.illinois.edu/rds/
[8] https://databank.illinois.edu

responses for each question, as well as any unique perspectives that emerged from the conversations.

Question 1: What are the major challenges facing RDMI today?

- **Training**: More researchers understand the "why" of data management, but require further training on the "how." Many of the best tools are still too difficult for the typical researcher to use without training. Also a concern about who will provide this training; currently libraries and research computing centers, but are there enough people?

- **Incentives**: Lots of "sticks" (e.g. funding agency or journal requirements) but very few "carrots" for sharing data. RDM is usually not considered as part of tenure and promotion process, so researchers aren't motivated to spent time on it. The sharing of unpublished data, particularly negative results, should also be encouraged.

- **Governance**: Who is responsible for research data management at the institutional level? Who sets the RDM policies and budgets for the entire campus?

- **Pace of Change**: The "data surge" is not over — as more researchers turn to sensors and other data-streaming technologies, data is growing faster than bandwidth and storage. Meanwhile, data management technologies, requirements from journals and funding agencies, and the data usage of research domains are constantly in flux.

- **Cost/Funding**: Research data management is restricted by a lack of effective cost models, funding that is often insufficient for proper research data management, and a lack of priority-setting from administration and leadership. The problem deepens as researchers consider longer-term storage — who will fund data management on a timescale of decades?

- **Compliance**: Institutions need expertise to interpret regulations and develop guidelines, including data use agreements. Disciplines need stronger and more consistent guidelines that can drive data management behavior and reduce uncertainty, while respecting that different fields and groups use data in different ways.

- **Data/metadata formats**: There is often no established consistency in data or metadata formats even within a single discipline or research topic, and no agreed-upon body to set those standards.

- **Security**: Many institutions are still struggling with how to properly store and share protected data — classified, health, or containing personal information — and remain compliant with complex regulations from multiple bodies. On the other end, many

research groups are unaware of security concerns with their data and do not handle it using proper procedures.

- **Curation**: Repositories should not be data dumps, they need to be organized and annotated. But who is responsible for this critical step?

Question 2: How are institutions addressing these challenges?

- **Changing IT Culture**: Make IT departments more flexible, responsive, and willing to help researchers use the tools they want to use.

- **Broad Solutions**: Funding agencies are stepping up to support training programs and tools for discoverability and sharing that make RDM easier for the typical researcher. Some of these grants fund "support services" from the library and IT, instead of researchers themselves.

- **Incentives**: Groups are starting to assign citable DOI numbers to datasets and software to encourage storage and publication. Some initiatives encourage curation by providing data storage, but require sharing if the researcher wants to access others' data.

- **Outreach**: Data librarians can attend both general and field-specific discussions to "spread the gospel" of RDM.

- **Hiring**: Bringing in more data librarians, archivists, research computing specialists, including dedicated staff for handling restricted data. However, there is still more work to do on building in appreciation and credit into these jobs, which are not traditionally recognized by science.

- **Funding Models**: Concentrate top-down funding on RDM services that are not valued by individual researchers but are valued by the institution; for example, providing a well-run secure service.

- **Defining Guidelines**: At the institutional level, gathering faculty with library and IT staff to discuss policies and determine infrastructure needs.

- **National Storage Options**: Research data services such as the upcoming NIH Data Commons[9] and the National Data Service[10], as well as domain-specific efforts such as

[9] https://commonfund.nih.gov/commons
[10] http://www.nationaldataservice.org

GenBank[11] and IRIS[12] are starting to provide off-campus resources for researchers to store and share data.

- **Open Scholarship**: Appeal to researcher's duty to conduct transparent and reproducible research by making data open access and properly connected to scholarly publications, framing RDM as part of the larger research landscape.

Question 3: How are these challenges impacting research and discovery today?

- **Slowing Discovery**: The difficulties of figuring out compliance, funding, and guidelines around research data management get in the way of time spent on research, and may even discourage innovative work. Additionally, poor data sharing infrastructure restricts the potential of big data research and collaborations between institutions.

- **Reduced Reproducibility**: Without widespread data publication, reproducibility is difficult, if not impossible. Coverage of this "reproducibility crisis" in science has also resulted in a decline in public trust of research.

- **Gaps Forming**: Younger researchers are more embedded in digital culture and more likely to practice good data management throughout their career, while older researchers want to archive their data but wait until the end of their career. Also, a separation is developing between large institutions that can afford advanced cyberinfrastructure and smaller schools who cannot make the investment.

- **Redundancy**: A lack of coordination between and within institutions and national standards leads to "reinventing the wheel" on RDM solutions, instead of using already established technologies and strategies.

Panel 2: Selected Examples of RDMI

As an implementations-focused workshop, an important portion of the agenda was presenting use cases of research data management efforts currently operating. The 2017 conference invited representatives from four national- and international-scale initiatives working to expand data preservation, publication, discovery, and reuse across institutions and disciplines. Each speaker presented their organization's mission, accomplishments, needs, and challenges thus far.

[11] https://www.ncbi.nlm.nih.gov/genbank/
[12] https://ds.iris.edu/ds/nodes/dmc/about/sections/data-management/

Moderator Rachana Ananthakrishnan, director of product management and design at **Globus**, described research data management implementations as "science accelerators," technology that transforms ad hoc storage solutions and isolated cyberinfrastructure into discoveries. While the software is already there to accomplish this goal, the trick is to make it easy for researchers to use these services. Ananthakrishnan provided example of Globus services used at Argonne National Laboratory's Advanced Photon Source and the National Center for Atmospheric Research data archives that enable researchers to move, store, share, and discover data.

Since 2013, the **Research Data Alliance (RDA)** has promoted data sharing by building "data bridges" to eliminate barriers across technologies, disciplines, and countries. Leslie McIntosh, the executive director of the RDA - US region, said their focus is on identifying problems with data, coming up with recommendations, spreading them to the scientific community, and soliciting feedback to improve those standards. The process prevents research groups from "reinventing the wheel" on processes such as citing evolving data; instead, they can take RDA standards and adapt them to their own practices.

> **"Research is hard. The technological advances, the increased data, the complex analytic techniques, they've all increased scientific discovery, yet they've convoluted the process of scientific work and reproducibility. It's made science harder while it's made some discoveries easier." - Leslie McIntosh**

McIntosh mentioned several needs for research data management, including support for the personnel who run data systems, increased investment in hiring research data specialists, and better metrics to measure both success and failure of technical solutions. An organization such as the RDA, which is dedicated to identifying the challenges on an international level, can help produce broad recommendations that will avoid "siloed" solutions that limit research.

But a domain-specific solution doesn't necessarily create new silos, as demonstrated by Tom Murphy, director of computing and network services for the **Inter-university Consortium for Political and Social Research (ICPSR)**. Established in 1962 and based at the University of Michigan, ICPSR maintains a data archive of more than 250,000 files in social and behavioral sciences. The focus, Murphy said, was not just on storing data, but ensuring that researchers have access to relevant data when they need it and creating value through data reuse.

As such, ICPSR provides a broad range of services across the research data life cycle, including data ingestion, curation, discovery tools, metadata editors, and more. Yet many of the tools built for their consortium's data were designed to handle data types from multiple disciplines, Murphy said, and can be adapted for use by other communities because of their commitment to open source.

The **National Data Service (NDS)** approaches this same interoperability from a domain-agnostic perspective, coordinating data stakeholders and incubating data projects for a variety

of different communities, including astronomy, biology, engineering, and material science. A consortium of universities, libraries, archives, publishers, and computing and data centers, NDS supports the publication, discovery, and reuse of data. Jim Myers, an associate research scientist at NDS, talked about how the consortium is shifting from an emphasis on projects and prototypes to an infrastructure perspective, given that a universe of "good enough" tools for data sharing, analysis, visualization, and computation have now emerged.

As an example of one such success story, Myers used SEAD (Sustainable Environment Actionable Data), a project initially funded by the National Science Foundation and now provided as a resource by the NDS. SEAD hosts over 4 terabytes of data used for dozens of publications in assorted fields and provides persistent identifiers for datasets so that authors receive citations when their data is used — an example of incentives driving good research data management practice.

Finally, the **Digital Preservation Network (DPN)** was introduced by chief technology officer Dave Pcolar. Emerging from the library and archive community rather than domain researchers, the DPN is a community-driven model dedicated to long-term preservation of the scholarly record, particularly in "dark storage" where it is not frequently accessed. Unlike the other more centralized efforts described in this session, DPN is a federated network of archives, with five nodes around the United States that ingest and replicate data, including provenance and chain of custody information.

Currently, DPN is piloting a new model of peer-to-peer preservation, which can draw upon institutional resources and repositories and further decentralize the network. DPN is piloting several different peer-to-peer models — including one-to-one, one-to-many, and many-to-many partnerships — and hopes that this new approach will encourage collaboration between partners to align and complement skill sets.

> **"Digital preservation isn't a single app by a single institution, it's a series of activities over time by a community." - Dave Pcolar**

In the discussion, panelists were asked about incentives that focus on the early stages of research, rather than just publication and sharing at the end of an experiment. Besides persistent identifiers for datasets, which researchers may not fully understand the value of until later, tools that stream data from laboratory instruments to office computers or automate metadata annotation can both make the research experience more efficient while in progress and promote good publication and preservation practices later.

Panel 3: RDMI and Industry

The first day's final panel invited representatives from the private sector to talk about research data management solutions they offer scientists and institutions. With one of the largest

academic publishers (Elsevier) and two of the world's leading technology corporations (Amazon and Microsoft) on the panel, the talks explained how many of the most buzzed-about tech innovations — including the cloud, edge computing, and deep learning — can help modernize research and fulfill the potential of data-driven science.

As research of all stripes grows more data-heavy, **Elsevier** has re-examined its publication model to help adapt its journals to new scientific practices. The traditional paper-based report of methods and results is no longer sufficient to adequately support reproducibility, transparency, and the communication of computational protocols. Thus, in addition to reimagining the format of journal articles and peer review processes, Elsevier is itself developing a suite of tools that aid in research data management to help researchers fulfill the new meaning and promise of scientific "publication" in the 21st century.

These tools, presented by VP of Research Data Collaborations Anita de Waard, include Hivebench[13], an electronic lab notebook to store workflows and protocols, Mendeley Data[14], a data repository for datasets up to 5 gigabytes in size, and DataSearch[15], which searches over a collection of repositories from Mendeley and other sources. Elsevier is also exploring new journal requirements — researchers must now include a citation or link to data in an article or make a statement as to why it cannot be shared — and publication models, such as data journals where the dataset itself is published with supporting information. Additionally, some Elsevier journals now require authors to submit software code used in the paper, and peer reviewers are asked to re-run the code on the data to validate results and claims.

Tools currently under development at Elsevier include a Mendeley data management platform that integrates many of the above tools with systems in use at academic institutions and new metrics for measuring data publication, citations, downloads, and other features of data sharing. In order to establish a baseline for these metrics, Elsevier conducted a survey of data sharing practices among researchers, which found high enthusiasm for sharing data (69% considered it important, 73% wanted access to other people's data) far outstripped confidence (only 37% believed there was credit for sharing, only 25% believed they were adequately trained to share data). Barriers cited by respondents included privacy concerns, ethical issues, and intellectual property rights, as well as confusion and variance over who oversees RDM at institutions.

At **Microsoft**, the mission for helping scientists is to "let researchers be researchers." The rising flood of data means a lot more work for researchers, many of whom are still running their laboratories largely on desktops and laptops, said Gaurav Hind, Cloud and Big Data Architect, US Education for Microsoft. To ease this strain and propel the shift towards a unified research community, new computational tools and campaigns are needed to fulfill the FAIR principles of findable, accessible, interoperable, and reusable data.

[13] https://www.elsevier.com/solutions/hivebench
[14] https://data.mendeley.com
[15] https://datasearch.elsevier.com/#/

One strategy Microsoft uses to work towards those goals is the creation of problem-specific data commons, bringing together researchers studying the same topic and asking them to decide what tools will help them collaborate. One example is the ALS Knowledge Network, a secure cloud where the community of ALS physicians and researchers can share genetic and clinical data to study the degenerative motor disease.

The company also continues to develop research tools, many connected to their Azure cloud platform and Cosmos DB database. In some cases, they've worked to make already popular tools operate on their own services, such as the ability to run Jupyter interactive Python notebooks on the cloud. For discoverability, the Microsoft Academic semantic search engine uses machine learning and natural language processing for deep exploration of research literature that goes beyond traditional keyword-based searches.

> **"The future is here, it's just not evenly distributed yet." - William Gibson, quoted by Sanjay Padhi**

In just 10 years, **Amazon Web Services** has grown from providing 80 different services to over 1,000 features and products used by government agencies, educational institutions, and the nonprofit sector for analytics in the cloud. Head of AWS Research Initiatives Sanjay Padhi focused on several different AWS capabilities that can help research groups, such as elasticity, edge-based computing, and image analysis.

Some of these services are used by extremely large collaborations, including the CMS detector at CERN, which produces 10 *petabytes* per second of data. Though the data is filtered down to "merely" 100 to 1,000 megabytes per second before it is distributed around the world for analysis and archiving, this filtration must remain flexible, capturing more data when there is an event of interest and less when there is not. AWS provides on-demand auto-expansion for the CERN cyberinfrastructure, giving them more compute cores when more data is collected.

AWS provides similar streaming and edge-based analysis services to websites, campuses, and cities building "smart" infrastructure for street lighting, intersection traffic control, and pothole detection. Their new Rekognition image analysis service, based on deep learning neural network models, can help research on early detection of diabetes and cancer in medical images, Padhi said.

Breakout Session 2: Best Practices and Benefits of RDMI - Impact on Science

After much of the day's panels discussed the persistent need for standards in RDMI, the workshop's second small group discussion focused on what best practices have been established, and how these help address the challenges identified in earlier sessions. Participants were also asked to end the first day on a positive note, brainstorming the most

significant benefits that research data management has provided so far in science. As with the earlier session, groups presented their conclusions to the full audience at the end of the discussion period. Common themes are listed below.

Question 1: What best practices have been identified in RDMI?

Question 2: How do these best practices address the major challenges identified earlier in the day?

Centralization & Collaboration: Centralization facilitates good infrastructure, and provides researchers with a single place to ask questions and receive training, avoiding confusion. It's also important to coordinate RDM efforts across campus and utilize the strengths of different entities, including IT services, libraries, research computing centers, and computational scientists.

Sharing resources and staff keeps costs down, leads to better and more consistent metadata management, and makes data and software products easier to reuse.

Ease of Use/Automation: The easier RDM tools and systems are to use, the more likely researchers are to use them early and often in the research process. Specialists should strive to meet researchers where they are most comfortable, whether that's at the command line or with browser-based tools. To support usability, it's important to collect frequent feedback from users and improve services accordingly.

These principles improve adoptability and compliance among researchers and increase usage throughout the research life cycle. Making RDM tools simple and automated also minimizes the need for training researchers in using complicated software, and prevents researchers from feeling overwhelmed by a multitude of best practices —they can be built into the tools.

Portability/Interoperability: Whenever possible, use open source software that can solve problems across disciplines and will work with other institutions. Similarly, interoperable metadata schemas should be used for data collections to enable discoverability, conversion, and exchange of data.

Following these practices improves the continuity of data, should it need to be ported to new systems in the future. Using accepted metadata schema also supports data curation and reusability by the original research group, an added incentive.

Hiring & Staffing: Hire data management professionals with expertise in science, IT, and library/curation (if you can find them). It's also critical to maintain discipline-specific curators that can handle the unique data challenges of individual fields. For schools that don't have the resources to hire this specialty staff, joining multi-institutional research computing groups may provide the necessary expertise.

Adding general data management staff helps with funding needs, as individual projects don't need to hire data professionals themselves, while hiring domain specialists alleviates problems caused by the lack of common vocabulary between fields. Identifying a need for dedicated data management staff can help influence institutional investment from administration.

Cost Sharing: There's a risk to the "freemium" model of offering a product or service for free initially, then charging when usage goes over a certain size. Some schools have started charging departments, rather than individual faculty, for storage. Others subsidize storage instead of offering it for free, using the retail price of hard drives to set the amount that faculty pay.

Providing clear and simple pricing helps researchers budget accordingly for data management, aiding compliance, and funding.

Training: Several different strategies were proposed to bring training to researchers at all stages of their career. Some institutions supplement live workshop-style training with online modules, to reach faculty that are less likely to attend events. Multiple groups suggested teaching graduate students the principles of RDM early in their education, and new faculty upon their arrival at campus.

Training in research data management can also serve as continuing education for best laboratory practices, including documentation, ethics, and reproducibility. These courses can also emphasize the benefits and incentives for researchers to properly store and share data. Providing younger researchers with more of an appreciation for computational science, and it will help change the culture long-term.

Follow Established Standards: For sensitive data including patient data, classified information, and data related to defense and arms, it's essential for RDM services to follow standards established by federal law.

Adherence to these standards, beyond its legal necessity, also creates secure infrastructure for all data, and teaches staff and researchers effective data management.

Data Management Plans: Work with researchers to create data management plans early in the research process, and include funding for its components in grant applications. Create plans with the end of the project in mind; ask researchers to think about what data outputs they will make available in the future.

Because funding agencies increasingly demand data management plans in grant applications, they offer an opportunity for improving data practices through training and help researchers request adequate funds for supporting data management.

Dataset Minting: Assign persistent identifiers such as DOI or ORCID numbers to datasets from the beginning so they can be cited later, which is both good practice and an added incentive for researchers.

Unique identifiers aid search and discoverability by other researchers, and make datasets a citable object, which can alleviate intellectual property concerns and help institutions evaluate faculty data practices as part of their tenure and promotion reviews.

Question 3: What significant benefits to science are being achieved?

- **Quality Control**: Metadata helps pick up problems in quality control and validation before they get to the publication stage. It also makes sure the data is interpretable by the community and doesn't need the researcher to explain.

- **New Science**: With the wider availability of data, scientists are asking different and new questions. Easier data sharing broadens research and encourages interdisciplinary and multi-institutional collaboration, such as large-scale genomic studies.

- **Improved Reproducibility**: The ability to validate scientific findings and data means there is less need to redo flawed research, leaving more funding for new science. The transparency of open data serves as a check and balance on research. Software is better documented and higher quality, making it easier for other groups to use.

- **Quicker Solutions**: With data management tools, researchers can spend more time doing their research, and less on repetitive tasks

- **Cost Savings**: Centralizing data services makes them more efficient and cost-effective for researchers, the institution, and even funding agencies, who can avoid funding the construction of infrastructure for each project.

- **Accessibility**: Data is increasingly available to parties outside of academic research, such as high school teachers, citizen scientists or people from other industries and fields.

- **Increase Research Value**: The "long tail" of research data enables studies on latent data that may otherwise have been discarded or lost. More longitudinal studies can also be conducted as older data is archived long-term.

- **Modernizing Scientific Communication**: Data-heavy research has provoked rethinking of scientific publications and the communication of science, including peer review, and the types of research outputs that should be made available.

- **Institutional Reorganization**: As demand for data and computation services increases among researchers, institutions are motivated to knit together services and programs in innovative ways to advance science

Panel 4: Funding Agencies Perspective and Plans for the Future

Throughout the first day of the workshop, panels and discussion groups frequently mentioned funding agencies as important drivers of research data management practices. As agencies such as the NIH and NSF have started requiring data management plans in grant proposals, academic support staff have seen increased demand for training, resources, and guidance among campus researchers. The workshop's final panel invited representatives from four national funding agencies to present their perspective, internally and externally, on data management and policies and how they expect them to further evolve in the near future.

Since 2011, the National Science Foundation has required data management plans as part of all funding proposals. The policy supports philosophy of data as "the core element of the work" and intrinsic to the nature and integrity of science, said Amy Friedlander, Deputy Directory of the NSF Office of Advanced Cyberinfrastructure. The challenge for the agency is to create coherent policies that encourage robust science, but still recognize the differences among the broad range of science and engineering disciplines that the NSF supports, including social, physical, mathematical and computer sciences. These policies are designed to protect the evidentiary base of research and enable communication of results, within legal and resource limitations.

Those principles have informed several new NSF directives over the last six years, including the data management plan requirement. A 2013 policy on data citation established that datasets are a first class research product, equivalent to peer-reviewed journal publications and subject to the same standards. The 2016 public access policy required that all NSF-funded journal articles and conference papers — and the underlying data — must be available for download, reading, and analysis free of charge.

The requirements for NSF data management plans reflect these policies, as researchers are required to detail the types of data a project will generate, the data standards they will use, and their plan for data access, sharing, archiving, and reuse. These plans are evaluated during the merit review process; if no plan is submitted, a justification statement for not needing data management is reviewed instead.

The agency continues to revise the structure of these data management plan requirements in collaboration with the research communities of the disparate disciplines they fund. The NSF is also examining how they can expand these policies to cover additional research products, such as software, and the role of cyberinfrastructure, including science portals, middleware services,

and on and off-campus resources that keeps data "in flight" instead of at rest in a depository (inspired by a LIGO paper).

> **"We don't want to just manage data -- we want to use and reuse data, and extract maximum value from it." - Jeff de la Beaujardiere**

The National Institutes of Health policies on data sharing and management remain more of a "moving target," said NIH research data informationist Lisa Federer. The 21st Century Cures Act[16], passed in late 2016, implemented measures to support data sharing among NIH researchers, but the exact NIH policies — including a data management and sharing plan, similar to the NSF requirement — have not yet gone into effect. One significant hurdle is the different considerations around sharing data in the context of clinical trials where sensitive patient data is involved. As a result, the NIH implemented and is further considering several different data policies, including separate standards for how and where data from clinical trials, human genomic research, and other areas should be shared.

In addition to policy, many efforts are underway at the NIH to support research data reuse, funded by the Big Data to Knowledge initiative[17]. One such tool is DataMed[18], a federated search across multiple data repositories to find datasets on different subject matters, akin to the PubMed service for searching journal articles. Another effort is Common Data Elements[19], which provides sets of metadata or data dictionaries that researchers can use to ensure that their datasets will be interoperable with national databases on specific subject areas.

Unlike the NSF and NIH, which predominantly fund and set policies for extramural science, the National Institute of Standards and Technologies (NIST) is focused primarily on in-house research. Their data management policies are derived from the 2013 White House Office of Science and Technology Policy (OSTP) directive[20] on making data publicly and freely available — with the exception of some standard reference data used by industry, for which NIST can collect fees to offset costs of collection and dissemination.

NIST Director of the Office of Data and Informatics Robert Hanisch discussed the agency's current initiatives to support and encourage data sharing, including the building of a new public data repository and search portal, assigning DOIs to all public data sets, and creating new global data discovery and annotation tools. NIST is also hosting a repository for the Materials Genome Initiative, and constructing a Laboratory Information Management System (LIMS) to automatically capture experiment metadata and move it into the proper database.

---

16 https://www.fda.gov/regulatoryinformation/lawsenforcedbyfda/significantamendmentstothefdcact/21stcenturycuresact/default.htm

17 https://commonfund.nih.gov/bd2k

18 https://datamed.org

19 https://www.nlm.nih.gov/cde/

20 https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Similarly, the National Oceanic and Atmospheric Administration (NOAA) is predominantly focused on making its own internally-collected data widely and freely available. The NOAA Data Catalog[21] currently holds over 70,000 datasets, including big data collected by satellites, autonomous vehicles, and human observers around the planet. To motivate this data publication, the agency attempts to automate as much of the data indexing process as possible for researchers, only assigns NOAA DOIs to data that is present in the archives, and requires data management plans for data curation and publication.

But Jeff de la Beaujardiere, Data Management Architect for NOAA, emphasized that data management in and of itself is not the end goal for the agency. If the ultimate purpose of preserving and publishing data is to extract maximum value from it, simply providing a huge dataset is not the ideal format. Instead, NOAA seeks to create a data services layer between the data and users, including decision-making tools that can condense large and complex data into actionable information.

Questions to the panel focused on matters such as data sharing policies for collaborations with commercial entities, conflicts between open sharing and sensitive data, the preservation and publication of software, new incentives to encourage data management practices, and how agencies can support training efforts. In their answers, all four funding agency representatives expressed optimism that researcher awareness of the need for data management — and advance planning for their research data needs — is improving.

Breakout Session 3: Future Directions for RDMI

The final session of the workshop focused on next steps and priorities for the future, based on the conversations of the previous day and a half. Because much of the previous breakout sessions focused on funding and sustainability, the groups for this discussion were asked to put those challenges aside and come up with additional problems and solutions that the community should focus upon going forward. The output of these conversations provided a blueprint for future meetings and collaborations to advance research data management acceptance and practices.

Question 1: In addition to the challenges regarding funding and sustainability, what are the top three priorities our community needs to address?

Question 2: How can we develop initiatives that will address the top priority identified above?

Funder Requirements: Research groups that receive funding from multiple sources often encounter mismatched data management requirements from their funders. There is a need for common policies and data elements.

[21] https://data.noaa.gov/datasetsearch/

Solutions: Funding agencies can relieve the burden on researchers by providing an extract-transform-load pipeline that automatically deposits data in the proper repository. For certain funding opportunities, it may make sense for agencies to predefine the RDM approach, focusing on existing resources, rather than ask researchers to provide their own plans and build new infrastructure from scratch.

Lack of Strategy (National, Domain, Campus): National infrastructure is needed for some domains, to enable data linkage and discoverability. At institutions, more administration leadership is needed to commit resources for additional hiring, training, and technology. Important decisions still need to be made at many schools and organizations about who is responsible for standards, infrastructure, compliance, and other important RDM considerations.

Solutions: Engage policymakers at institutions and organizations by hosting sessions including researchers, librarians, data managers, CIOs, and other administration staff to discuss institutional standards and resources. These sessions could be mediated by national organizations such as the Coalition for Academic Scientific Computation (CASC).

Curation: As more and more data is produced, communities are needed to curate these datasets as they go into repositories and research computing facilities. Currently, most of this task falls to libraries, but more trained staff and domain expertise is needed.

Solutions: Distributed data curation networks[22] can share the responsibility and combine skills in different areas.

Training: Researchers, students, and data professionals need more education in research data management practices and tools.

Solutions: Select faculty "champions" that will talk to their departments about the benefits of data management and guide them to training resources. Integrate data management training into graduate and undergraduate level science curricula, as part of the standard scientific process. Training providers such as Software Carpentry[23] can also be engaged to improve data and computational literacy among faculty and student researchers.

Automation/Integration: More tools embedded into the research environment are needed to help promote research data management without adding additional time and effort for researchers. Automated collection of metadata, in particular, will help enable reproducibility and reuse.

Solutions: Conduct more independent assessments of currently available tools and hear from end users on what works best for their research process, identifying the easy-to-use and effective solutions that should be used by the broader community.

[22] https://sites.google.com/site/datacurationnetwork/
[23] https://software-carpentry.org

Incentives/Enforcement: Despite new policies and requirements for sharing and publishing data, several holes remain. In many cases, the publication of scientific code and software is still not required by journals or funding agencies and the review process for data management plans lacks teeth. On the incentives side, tenure and promotion reviews rarely credit data management.

Solutions: A neutral party such as the Research Data Alliance could develop recommended incentives for governments and academic institutions, working with different disciplines to find appropriate standards.

Discoverability: Usage of contextual metadata must improve to support discoverability for both researchers reusing their own data and external groups finding and building upon completed work.

Solutions: Develop a "W3C" for data, to establish common standards that allow interoperability of data infrastructures and create a "national data inventory" that helps researchers find the data they need.

Question 3: Separate from the first two questions, should we continue the conversations started at this workshop, and how?

Many groups agreed with the idea that sessions or workshops dedicated to research data management should take place at domain-specific meetings, led by teams of people already familiar with the RDM world (e.g. computing staff, librarians, funding agency representatives). These programs would avoid "preaching to the converted," and increase awareness and best practices among new communities of researchers. Other outreach suggestions included writing op-ed letters to the *Chronicle of Higher Education* and other publications with academic audiences, promoting useful tools, researcher success stories, and the importance of research data management for knowledge sharing and scientific reproducibility.

Other groups sought solutions to keep conversations from the present workshop going, with technical suggestions including a bulletin board, a dedicated stack exchange, a Slack channel, or a mailing list. Another proposal suggested quarterly or bi-annual webinars addressing different areas of research data management, highlight topics of interest to both the RDM community at faculty researchers interested in learning more on the subject.

Conclusion

Over the course of the two-day workshop, several prominent themes emerged around the current status of research data management implementations and future challenges and opportunities.

Findings

- Researcher usage of RDM is increasing, driven by greater awareness of its importance among younger scientists, funding agency requirements for a data management plan in grant proposals, widely available and user-friendly tools, and the overall shift toward data-driven research in most research fields.

- However, as demand for these services increases, critical gaps remain in funding, staffing of trained research data professionals, and training of researchers in data management skills.

- The rapid expansion of the research data management tool ecosystem continues, with better solutions for data storage, sharing, and discovery emerging on a regular basis. Now, careful scientific assessments are needed to identify those that work best for users, and those tools should then be more widely promoted.

- Reproducibility is a key tenet of the scientific method, and research data management is crucial for reproducibility in modern science. Publication and sharing of data and other computational research products (e.g. software, workflows) enable the verification of research results.

- Libraries increasingly play an important role in research data management on campus, providing technical training, curation services, and assistance with assembling data management plans for grant applications. Whenever possible, campuses should coordinate RDM services and strategy between IT services, libraries, computing centers, and research faculty, as it avoids redundant spending and effort and utilizes expertise across the institution.

- Data management plans are quickly becoming mandatory for grant proposals to federal funding agencies. NSF created its data management plan in 2011, the NIH is currently reviewing policies to make these plans mandatory, and NOAA and NIST both require plans of some form for their intramural research. These requirements have driven up participation in research data management; however, the format of these data management plans and the standards for their peer review remain in flux.

- Businesses such as Elsevier, Amazon, and Microsoft are focusing their research data management efforts on creating tools for data discovery and sharing, creating new platforms, collaborative data commons, and elastic edge computing technologies.

- There is an intense need for data standards to ensure portability, interoperability, and easier data discovery. Even within single disciplines, consensus rarely exists on

standards for data and metadata formats, and there is no agreed-upon authority to set those standards.

Recommendations

- At the campus level, more administrations need to commit to supporting RDM resources and determine a sustainable strategy for long-term data storage, curation, and sharing. These conversations should involve all campus stakeholders (IT, library, researchers, computing centers) and determine governance, policies, and budgets for campus data management.

- In addition to research data management implementation workshops such as this one, create more RDM workshops at discipline-specific meetings, run by people with experience in data management but aimed at researchers who are new to the topic.

- Incorporate research data management training into all levels of curricula to reach new researchers and educate them that data management is an essential aspect of modern science.

- To appeal to senior faculty, find campus "champions" who can speak to the benefits of data management and help steer researchers to the proper resources at the institution. Also consider different formats of training, including webinars or online modules, to draw in researchers less likely to attend an in-person workshop.

- Create new incentives for data management at multiple levels. At academic institutions, consider dataset publication and sharing as part of the tenure and promotion process. National-level collaborations or agencies can encourage sharing by requiring it if a researcher wants access to others' data.

- Developers of research data management tools should focus on automation, "hands-free" functions that handle data during the natural workflow of science, instead of adding effort and time. These features increase adoption and compliance with data management standards, while also reducing cost and training demands.

- Hire additional data management professionals with expertise in science, IT, library curation -- if you can find them. Domain-specific curators are also important, but schools lacking resources to hire experts in each domain can form partnerships with other institutions to share expertise.

Overall, the workshop celebrated the progress made in the six years since the 2011 research data lifecycle management (RDLM) meeting in Princeton which established this community. More researchers than ever are interested in research data management, see its value for

scientific discovery and reproducibility, and have access to better and better tools for storage, sharing, and discovery.

However, alongside these success come new challenges. Increased demand for data storage and increasingly data-heavy research instrumentation and methods create new concerns around funding and sustainability. New demand for training and resources have strained staffing levels at many institutions, and the lack of a top-down strategy at many institutions for hiring and building infrastructure has slowed data-driven discovery and archival efforts.

But there are exciting challenges ahead as well. With more data management tools available than ever, researchers can start moving past the custodial phase of data storage and publication, and into the new potential unlocked by advanced data sharing and discovery functions. At the workshop, multiple examples of new collaborations within and across disciplines, fueled by the exchange of data, provided inspiration for continuing the important work of implementing research data management. As put by NOAA's Jeff de la Beaujardiere:

> **"We don't want to just manage data — we want to use and reuse data, and extract maximum value from it."**

Acknowledgements

The organizing committee for the RDMI 2017 workshop consisted of: committee chair Hakizumwami Birali Runesha, University of Chicago; Thomas Furlani, University of Buffalo; Ruth Marinshaw, Stanford University; Benjamin P. Rogers, University of Iowa; Wendy A. Kozlowski, Cornell University; Rachel Vincent-Finley, Southern University and A & M College.

We thank Lisa Arafune of CASC and Kimberly Grasch of the University of Chicago for their support in organizing the workshop.

Appendix 1: Position and Experience Papers

2017

1. Allan, G., Erdmann, J., Gustafson, A., Joglekar, A., Milligan, M., Onsongo, G., Pamulaparthy, K., Pardey, P., Prather, T., Senay, S., Silverstein, K., Wilgenbusch, J., Zhang, Y., and Zhou, P. (2017). G.E.M.S: An Innovative Agroinformatics Data Discovery and Analysis Platform. University of Minnesota.
2. Ananthakrishnan, R., Chard, K., Foster, I. (2017). Design patterns for data-driven research acceleration. University of Chicago and Argonne National Laboratory.
3. Bose, R. and Scap, B. (2017). Research Data Storage for Columbia's Zuckerman Mind Brain Behavior Institute. Columbia University.
4. Chourasia, A., Nadeau, D.R., and Norman, M.L. (2017). SeedMe: Data Sharing Platform for the research community. University of California San Diego.
5. de Waard, A. (2017). The Mendeley Data Management Platform: Research Data Management From A Publisher's Perspective. Elsevier.
6. Fearon, D.S. and Boehm, R. (2017). Addressing Gaps in Data Archiving for Large-Scale Computing Research at the Academic Research Institution. John Hopkins University.
7. Huffman, K. Hoogendyk, T., and Brainerd, E. (2017). XMAPortal: Data Management and Data Sharing for Motion Analysis Research. Brown University.
8. Johnson, A. (2017). Towards an Outcomes-Based Approach to RDM: Experiences from the CU Bolder Center for Research Data and Digital Scholarship. University of Colorado Boulder.
9. Myers, J. (2017). Where's My Universal Data Browser?. University of Michigan.
10. Runesha, H.B. (2017). DaLI: a Data Lifecycle Instrument for management and sharing of data from experiements and observations. University of Chicago.
11. Sierakowski, A.J. (2017). Craedl: Collaborative Research Administration Environment and Data Library. John Hopkins University.
12. Song, C., Smith, P., and Younts, A. (2017). Creating a seamless campus cyberinfrastructure to support data-driven domain science. Purdue University.
13. Wittenberg, J. (2017). Research Data Management in Context: Embedding Research Data in Open Scholarship at Indiana University. Indiana University.

Copies of these papers can be downloaded from https://rdmi.uchicago.edu/index.php.

2013

1. Adamus, T., Cheetham, J., Salo, D., Schryver R., and Wolf, A. (2013). Campus Level Research Data Management Services at the University of Wisconsin-Madison: Past, Present, Future. University of Wisconsin - Madison.
2. Agnew, G. (2013). Video Mosiac Collaborative. Rutgers University.
3. Barker, M. and Reed, M. S.C. (2013). A Research Environment for High Risk Data. Research Computing at University of North Carolina at Chapel Hill.

4. Basu Ray, J. and Broadway, R. (2013). RDMI at Dillard University - our "now" and "in future" strategies. Dillard University.
5. Bielefeld, R. and Warfe M. (2013). FISMA Compliant Research Environment at CWRU. Case Western Reserve University.
6. Bose, R. and Nurnberger, A. (2013). Columbia's Evolving Research Data Storage Strategy. Columbia University.
7. Chen, P. (2013). A High-Performance Computing Environment to Support Research and Teaching at a Minorirty Serving Institution. University of Houston-Downtown.
8. Crane, G. (2013). The SURA/ASERL Research Data Management Collaboration. Southeastern Universities Research Association.
9. Evrard, A., Westbrooks, E., and Broude Geva, S. (2013). Collaborative Efforts in Data Management at Michigan. University of Michigan.
10. Foster, I., Runesha, H.B., and Vasiliadis, V. (2013). Campus support for research data management: A perspective from the University of Chicago. The University of Chicago.
11. Freeland, C. (2013). Leveraging open access storage portals for the management of text-based research data. Washington University.
12. Greenberg, J., Rowell, C., Rajavi, K., Conway, M., and Lander, H. (2013). HIVEing Across U.S. DataNets. Metadata Research Center at University of North Carolina at Chapel Hill and DataNet Federation Consortium/Data Bridge: Renaissance Computing Institute.
13. Gregurick, S. (2013). DOE Data Management Systems and Biology Knowledge Base. Department of Energy.
14. Gurkan, D., Reilly, M., and de la Cruz Guiterrez, M. (2013). Research Metadata Exchange Design at University of Houston. University of Houston.
15. Hedstrom, M., Alter, G., Kouper, I., Kumar, P., McDonald, R.H., Myers, J., and Plale, B. (2013). SEAD: An Integrated Infrastructure to Support Data Stewardship in Sustainability Science. University of Michigan, Indiana University, University of Illinois at Urbana-Champaign, and Rensselaer Polytechnic Institute.
16. Huffman, K., Brainerd, E., and Combariza, J. (2013). The XMA portal. A web environment for data management. Brown University.
17. Ismail, S. and Tsung, S. (2013). Data, Data Everywhere, but Not a Byte to Share: Experiences from the Trenches in Building, Managing, and Supporting Faculty's e-Research Needs. Georgetown University Libraries.
18. Jack, M. (2013). High-Performance Computing at Florida A&M University. Florida A&M University.
19. Li, Y., Bern, P., and Carrier, T. (2013). The Library Data Services Development: Syracuse University Example. Syracuse University.
20. Matthews, A., Bobovych, S., Johnston, W., Banerjee, N., Cothren, J., and Parkerson, J. (2013). Map Dissemination in Disaster Scenarios. University of Maryland and University of Arkansas.
21. Molnar, P. and Kossorla, M. (2013). Data Management Position Paper. Clark Atlanta University.
22. Moore, R. (2013). Distributed Data Management Concepts. University of North Carolina at Chapel Hill.

23. Qin, J. and D'Ignazio, J. (2013). Filling the Gap between Project-Level Data Managment Needs and Disciplinary Data Repositories. Syracuse University.
24. Rohrs, L., Conte J., and Mistry, H. (2013). Data Management Planning and Services around the Data Life-Cycle at New York University. New York University.
25. Salayandia, L., Gates. A. Q., and Pennington, D. (2013). MetaShare: Constructing Actionable Data Management Plans through Formal Semantics. University of Texas at El Paso.
26. Strasser, C. and Cruse, P. (2013). The DMPTool and DataUp: Helping Researchers Manage, Archive, and Share their Data. University of California Curation Center, California Digital Library.
27. Underwood, G. and Marciano, R. (2013). Services Needed for Managment, Preservation and Access to Digital Records of Scientific & Engineering Research. Georgia Tech Research Institute and University of North Carolina.
28. Van Tuyl, S. (2013). Developing a Research Data Management Services Infrastructure at Carnegie Mellon University. Carnegie Mellon University.
29. Warner, B., Faerman, M., Newman, L., Wohlever, K., Evans, G., and Shah, P. (2013). Data Sharing in Ohio: A Discussion. Ohio Supercomputing Center.
30. Xie, Z. (2013). Facilitate Cross-Repository Big Data Discovery and Reuse. Virginia Tech University.

Copies of these papers can be downloaded from http://rdmi.uchicago.edu.

Appendix 2: RDMI Program

RESEARCH DATA MANAGEMENT IMPLEMENTATIONS WORKSHOP

Westin Arlington Gateway, Ballroom C, Arlington, VA

Day 1: September 14, 2017

8:00 AM	**Registration and Continental Breakfast**
8:30 AM	**Welcome: Overview and Goals** ***H. Birali Runesha***, AVP for Research Computing Director, Research Computing Center, University of Chicago RDMI Committee Chair
8:35 AM	**Introduction** ***Amy Walton***, Program Director for the Office of Advanced Cyberinfrastructure, NSF
8:45 AM	**Keynote: "Research Data Management Implementations: Towards the Reproducibility of Science"** ***Victoria Stodden***, Associate Professor, School of Information Sciences, UIUC
9:30 AM	**Panel 1: Campus Support & Challenges with RDMI** **Moderator: Rajendra Bose**, Columbia University, CASC Chair ***Juan de Pablo*** Liew Family Professor of Molecular Engineering, University of Chicago ***Preston Smith*** Director, Research Services and Support, Purdue University ***Rachel Vincent-Finley*** Assoc. Dean for Academic Affairs, College of Sciences and Engineering Southern University and A&M College ***Christie Ann Wiley*** Engineering Librarian, University of Illinois at Urbana Champaign
10:30 AM	**BREAK - Coffee and Tea**
11:00 AM	**Summary of Position and Experience Papers** ***Ruth Marinshaw***, Chief Technology Officer – Research Computing,Stanford University
11:30 AM	**Breakout Session – Small Group 1: Today's Challenges in RDMI**
12:30 PM	**Lunch - Ballroom D/E**
1:30 PM	**Breakout Session Report**
2:00 PM	**Selected Examples of RDMI** **Moderator: Rachanna Ananthakrishnan,** Director of Product Mgmt and Design Globus, Argonne ***Leslie McIntosh*** Executive Director, Research Data Alliance ***Dave Pcolar*** Chief Technology Officer, Digital Preservation Network ***Tom Murphy*** Director of Computing & Network Services Inter-University Consortium for Political and Social Research ***Jim Myers*** Associate Research Scientist, National Data Service

Day 1: September 14, 2017 (continued)

3:00 PM	**RDMI and Industry** ***Moderator: Benjamin Rogers,*** Director of Research Services, University of Iowa ***Anita de Waard*** VP Research Data Collaborations, Elsevier ***Gaurav Hind*** Cloud & Big Data Architect, US Education, Microsoft ***Sanjay Padhi*** Head of AWS Research Initiatives, Amazon AWS
4:00 PM	**BREAK - Light Refreshments**
4:15 PM	**Breakout Session - Small Group 2: Best Practices and Benefits of RDMI - Impact on Science**
5:15 PM	**Breakout Session Reports**
6:30 PM	**Cocktail Reception - Ballroom D/E**
7:00 PM	**Dinner - Ballroom D/E**

Day 2: September 15, 2017

8:00 AM	**Continental Breakfast - Ballroom D/E**
8:30 AM	**Agenda Review** ***H. Birali Runesha***
8:45 AM	**Panel: "Funding Agencies Perspective and Plans for the Future"** **Moderator: James Wilgenbusch**, Assoc. Director, Minnesota Supercomputing Institute, University of Minnesota ***Amy Friedlander*** Deputy Director for the NSF Office of Advanced Cyberinfrastructure, National Science Foundation ***Robert Hanisch*** Director, Office of Data and Informatics, National Institute of Standards and Technology ***Jeff de La Beaujardiere*** NOAA Data Management Architect, National Oceanic and Atmospheric Administration ***Lisa Federer*** Research Data Informationist, National Institutes of Health Library
10:15 AM	**BREAK** - Coffee and Tea
10:30 AM	**Breakout Session – Small Group 3: Future Directions for RDMI**
11:30 AM	**Breakout Session Reports**
12:00 PM	**Open Discussion and Wrap -Up**