National Science Foundation
WHERE DISCOVERIES BEGIN

Final Report:

Workshops to Gauge the Impact of Requirements for Public Access to Data Produced by NSF-funded Research in Mathematics and the Physical Sciences[1,2]

[1] Editors: Robert Hanisch (NIST), Michael Hildreth (Notre Dame), Leah McEwen (Cornell), Victoria Stodden (UIUC), Gordon Watts (UW), Daniel S. Katz (UIUC), Natalie Meyers (Notre Dame), Ashley E. Sands (UCLA)

[2] Funded by NSF-PHY-1457413

Table of Contents

Overview

It has become technically feasible to make public the research data derived from most projects supported by federal funds. Making research data broadly available opens the scientific enterprise to other experts and citizen scientists, enhancing the potential for further discovery and real world applications. As an additional benefit, an effective implementation of open access to data, interpretive tools, and results could be an important ingredient in restoring public trust in the scientific enterprise. Other benefits include ease of access, global reach, reproducibility of results, and reuse and repurposing of data, which will maximize investment in research.

In 2013, the White House Office of Science and Technology Policy directed federal agencies with over $100 million in annual Research and Development expenditures to develop implementation plans for expanding public access to the results of federally-funded research, including scientific data in digital formats[3]. The National Science Foundation's implementation plan, "Today's Data, Tomorrow's Discoveries" (NSF 15-52)[4], was released in March 2015. The report outlines a staged strategy for increasing public access to research results, beginning with enabling access to peer-reviewed journal articles and conference proceedings. It leaves in place the current Data Management Plan (DMP) policies related to the access and preservation of research data, but includes a clear statement of intent toward further openness. Modifications in policy, however, will be developed through dialogues with the research communities that would be affected: "Changes in the system that may result in guidance associated with DMPs will take place incrementally after consultation with the research community and will be implemented no earlier than FY 2016[5] ."

This report is a direct result of consultation with the research communities funded by the Mathematical and Physical Sciences (MPS) Directorate at the National Science Foundation (NSF). The goal of this effort is to provide feedback to NSF on current best practices with regard to research data curation, discovery, access, preservation, and re-use, and suggestions for areas of improvement and investment that could facilitate broader curation of, access to, and re-use of research data in the future. Sponsored by an NSF award (NSF PHY 1457413), the consultation process consisted of two workshops and extensive community outreach. The first of the two workshops was held in Arlington, VA, on November 19 and 20, 2015. Attendees represented the various MPS domains, and also included archivists, librarians, publishers, and computer scientists. The ideas that emerged from the first workshop were captured in a draft report shared with the broader community of MPS-funded researchers. Outreach efforts (described below) included presentations at several professional society meetings and a survey of authors whose work is published in American Physical Society (APS) journals. A second workshop was convened in December 2016

[3] http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[4] https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf

[5] ibid, p. 9.

to accept and respond to the input collected following the release of the draft report, and to produce this final report.

Executive Summary

This report contains a broad discussion of open access to data and what this might imply for researchers funded by the NSF's MPS directorate. The potential benefits of public access and sharing of data are recognized by many segments of the research community. However, during the deliberations encompassed by this project a broad consensus emerged that *the MPS research culture, data management tools, and archives infrastructures are not ready at this time for a move to requiring public access to all MPS research data*. One concrete step that the MPS community is willing to take at this time is represented by the following statement:

Data and other digital artifacts upon which publications are based should be made publicly available in a digital, machine-readable format, and persistently linked to those relevant publications.

This is the first in a series of steps outlined in this report as a roadmap toward achieving the *potential* future goal of making research data reusable. An implementation of this step would already represent significant progress toward that goal and would demonstrate a commitment to public access on the part of the MPS community.

A second conclusion that can be drawn from the discussions of this project is that, unsurprisingly, *different disciplines have a wide variety of current practices and expectations for appropriate levels of the sharing of data and other scientific results*. A discipline-specific policy discussion will be required to determine an appropriate level of preservation and re-use.

A third conclusion is that *the provision of public access to data entails costs in infrastructure and human effort, and that some types of data may be impractical to archive, annotate, and share. Cost-benefit analyses should be conducted in order to set the level of expectations for the researcher, his or her institution, and the funding agency.*

Finally, there was a consensus that *creating incentives toward the sharing of data is the primary way to accomplish broad adoption of open access to data as the norm.* Activities related to the preservation of data and other research products for public access are not sufficiently recognized and encouraged within current reward structures, nor are there sufficient tools to make data preservation and sharing easy. Incentives can come in many forms, from changes in hiring and tenure practices to guidelines or requirements by publishers and funding agencies. Combining many different forms of inducement will be necessary to move entire research communities towards widespread data sharing.

The four conclusions presented above represent the main concerns of the MPS community regarding issues of public access to research data at this time. Rather than conclude at this point, however, the workshop participants wished to present a forward-looking vision of what an open science future might look like, and suggest some strategies for attaining this vision. The future envisioned here is one where re-use of research data is the norm, and the necessary tools and infrastructure are in place to maximize the scientific potential of research data by sharing it with others. Whether or not this vision will or even can become a reality, many elements of a data open-access infrastructure are being contemplated or are under construction. Since it is not clear that policy-makers and those driving the efforts towards public access appreciate the full complexity of what is required to make shared research data useful, as trends toward public access to research data accelerate it will become increasingly important for funding agencies and researchers alike to appreciate the efforts, technological as well as sociological, required to reach such a future. These are the issues addressed in the second part of the report. By including these here, we hope to explore this complexity from the perspective of researchers in the MPS community.

After a definition of terms and a survey of best practices, which have bearing on the current research environment, the report discusses the appropriate elements that would characterize a functional open access data infrastructure. These include a description of the components of a research project that should be stored along with and linked to the data and publication to enable re-use of the data to derive new scientific results:

- **Application software:** the software used to capture or create, process, and analyze the data.
- **Workflow:** instructions, frameworks, scripts, or other high-level code used to capture data and metadata and to run the application software.
- **Software environment:** a specification or instantiation of the requisite low-level software and hardware, including operating system, architecture, libraries, machine state, etc., that are necessary to run the application software and workflows.
- **Simulation capabilities:** the capability to run the application software with different parameters than used to generate the original data.
- **Documentation:** a description of the application software, workflows, standards, and other information describing how the data were created, derived, processed, and analyzed, and validated.
- **Data characterization:** documentation of the data themselves (formats, content, provenance, etc.) and the metadata that describes them and makes them discoverable and re-usable.
- **Measurement parameters:** sample metadata, instrument metadata, measurement techniques, calibration, reference standards, etc.

Different levels of re-usability can be characterized by a lesser or greater reliance on these additional pieces of information. Also discussed in the report is a set of technical

attributes and requirements for the archive and access infrastructure. Currently, no cyberinfrastructure system exists at scale that satisfies all of these requirements. However, we suggest a series of pilot projects that could supply a set of building blocks for the modular development of this infrastructure. Finally, there is an exploration of which cultural norms would need to be shifted for broad acceptance of public access to data as the new paradigm. We suggest that this issue not be overlooked; willing participation of the research community will be critical for the success of a move toward open data and, perhaps, open science as the norm. Finally, we provide suggestions for ways that the NSF can accelerate the evolution of the open access[6] landscape. In addition to funding various pilot projects that can provide critical pieces of open access infrastructure, the NSF could improve communication and set ground rules via the following suggestions:

- The MPS community needs examples of excellence in order to lower the barriers to preparing scientific results for sharing and archiving. NSF could highlight and disseminate best practices for data preservation and sharing. This could take the form of encouraging publication of excellent examples of Data Management Plans, showcasing the state of the art in published datasets with discoverable products, highlighting important scientific results derived from re-use of public datasets, etc. Having the DMPs linked to abstracts and the data or other products resulting from the grant would be an excellent resource for researchers seeking to follow these exemplars.
- The MPS community needs guidelines for trusted repositories in line with the NSF policy on disseminating research[7]. NSF does not have the infrastructure to store all of the research data that scientists may want to archive. Researchers have available to them a variety of repositories of varying quality and sophistication, not all of which meet ISO 16363 standards for trusted repositories. Some minimum set of requirements should be established so that MPS scientists know that they have stored their data in a repository that meets the standards for a trusted repository as far as NSF policy is concerned. These guidelines should address, at a minimum, such concerns as data security, licensing, and the quality of bit-level integrity checking.

Synopsis of NSF Public Access Policy

For informational purposes, the NSF Public Access policy presented in NSF 15-52[8] is summarized here.

---

[6] Note: in this document, we use the terms “Open Access” and “Public Access” more or less interchangeably. That is, “public” results are, by definition, open to anyone who can access them. No cost or payment structure is implied, i.e., public results may not necessarily be accessed free of charge. In the latter portions of this report, we use “open access” to suggest a more organized and coherent data and knowledge preservation structure.

[7] https://www.nsf.gov/pubs/manuals/gpm05_131/gpm7.jsp - 734

[8] https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf

NSF 15-52 states: “This plan sets forth a framework for increasing access to the results of NSF-funded research and leverages existing NSF policies that provide for data sharing, data management plans, and evaluation, monitoring, and compliance. NSF will continue to identify additional approaches, involving public and private sector entities, and will continue efforts to improve public access to research data. NSF will explore, along with other agencies, how best to achieve improved public access, including data storage and preservation, discoverability, and reuse with a particular focus on data underlying the conclusions of peer-reviewed scientific publications resulting from federally funded scientific research.”

Quoting from 15-52: “NSF’s data-sharing policy states: ‘Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing’ (PAPPG’s *Award & Administration Guide, Chapter VI.D.4*).” The main question facing the community is how to facilitate this sharing to maximize benefit while minimizing cost (both effort and time) to the researcher.

All funded grant applications submitted after January 2016 are required to deposit the “version of record” or the final accepted peer-reviewed manuscript of any resulting journal publication or paper in juried conference proceedings or transactions into a public access compliant repository designated by NSF, initially the NSF archive hosted by the Department of Energy PAGES system and accessed through grants.gov. In addition, “Data and associated outcomes that result from NSF-funded research and are subject to the existing DMP requirement” are considered “in scope” and subject to whatever policies are issued in the first round of considerations of open access.

NSF 15-52 also lays out several future directions for NSF policy considering research results and the underlying data. For example,

- “NSF will explore whether all data underlying published findings can be made available at the time of publication.“
- “NSF expects to explore a series of options to leverage existing data repositories, extend approaches already in use in the development of DMPs, develop standards for repositories and metadata in consultation with the community, and enhance reporting and evaluation procedures.”
- With regard to depositing research data: “Over the next several years, NSF will consult with the research communities to develop discipline- specific guidance and best practices.” Until new guidelines are announced, the handling of data is left to the practices outlined in the Data Management Plan submitted with each proposal.

Introduction

As mentioned above, the purpose of this report is to present the views of the

Mathematics and Physical Sciences (MPS) scientific communities on the sharing of data. As such, it represents an important step in the "consultation of community" that is a major component of the discussion outlined in NSF 15-52. Two workshops, held in November 2015 and December 2016, supported a wide discussion of various issues related to the sharing of research data, including current practices and problems, anticipated difficulties, and "best imagined" scenarios. There was broad participation from all MPS constituencies, as well as archivists, librarians, computer scientists, and guests from the publishing sphere. Experts from NASA, NIST, NIH, and DOE who are currently engaged in formulating their own policies on data sharing also attended.

In the course of discussion, several broad themes emerged. Most importantly, and unsurprisingly, individual disciplines within MPS have very different approaches to data processing, data analysis, and data sharing. This suggests that it will be very difficult to implement a "one size fits all" solution to open data access and storage. Storage solutions, including the metadata used to search for and retrieve data, will likely need to be domain-specific so that they can meet the needs of the individual research communities they hope to serve. A second important theme that arose was the distinction between the needs of what is often termed as the "long tail" of small projects compared with those of "big" or modestly large science projects. While large projects may be able to invest in developing the tools necessary for data archiving and sharing, the individual researcher has more difficulty and less motivation for doing so; these differences should factor into the recommendations such that they are appropriate for groups or projects at a variety of scales.

Thirdly, there was general agreement that presenting incentives toward the open sharing of data is the primary way to accomplish broad adoption of open access to data as the norm. These incentives could take a number of different forms. For example, the tools used to capture the necessary provenance information about a dataset and its processing could make scientific workflows easier, thus making it an advantage to work in a manner that also leads to preservation and sharing of data. A second example might be the creation of a different reward structure, so that either additional funding or another advantage is provided to those who prepare and share re-useable data. Along this line, giving appropriate publication credit to acknowledge the work required to prepare and curate new datasets may also spur these efforts. If data citation standards, e.g. those from FORCE11[9], were widely followed, one might be able to use the number of citations of a given dataset to provide some basis for an evolved reward structure. Similarly, the software citation principles[10] produced by another FORCE11 working group[11] provide the foundations on which software could be cited and indexed, and those indices could then be used as part of the reward structure.

---

[9] https://www.force11.org/group/joint-declaration-data-citation-principles-final, https://www.force11.org/node/7784

[10] Smith AM, Katz DS, Niemeyer KE, FORCE11 Software Citation Working Group. (2016) Software citation principles. PeerJ Computer Science 2:e86 https://doi.org/10.7717/peerj-cs.86

[11] https://www.force11.org/group/software-citation-working-group

In addition to incentives, mandates are starting to appear, both from publishers and funders. As a funder example, the Wellcome Trust updated their policies for their grantees on 10 July 2017 to include, "As a minimum, the data underpinning research papers should be made available to other researchers at the time of publication, as well as any original software that is required to view datasets or to replicate analyses."[12] In addition, the EU mandates for the Horizon2020 funding initiative that all publications be publically accessible[13], and encourages all researchers to deposit their research data in a manner that would make it reusable[14], following FAIR principles[15]. As an example from a publisher, PLOS states, "PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception."[16]

These broad themes ran throughout the discussion of current practices and potential future activities with regard to the curation and sharing of research data, and their implications should be considered when discussing any future plans.

Stakeholders in the open sharing of MPS data include a) funding agencies and foundations, b) research institutions and libraries, c) non-profit and commercial publishers, d) data and software repositories, e) researchers, f) instrument and software vendors, g) scholarly societies and h) other national and international scientific bodies, including industry. Existing and evolving roles among stakeholders overlap and, in some cases, compete for community adoption of standards, protocols, and workflows. To the extent that broader entities involved in coordination and interoperability (e.g., CODATA, Cross Ref/Mark, DataCite, FORCE11, ORCID, RDA, etc.) provide direction by engaging all stakeholders in demonstrating the benefits of sharing data, there is likely to be more robust resolution of areas of potential conflict and opportunities for further collaboration around common problems. This can be challenging and complex, given the heterogeneity of MPS data across disciplines.

Levels of Data Curation and Sharing

A critical consideration in any discussion of data sharing is deciding what kinds of data are to be shared[17]. This establishes the expectations for the level of potential re-use

---

[12] https://wellcome.ac.uk/funding/managing-grant/policy-data-software-materials-management-and-sharing

[13] ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

[14] ec.europa.eu/research/participants/data/ref/h2020/other/hi/oa-pilot/h2020-hi-erc-oa-guide_en.pdf

[15] http://datafairport.org/fair-principles-living-document-menu

[16] http://journals.plos.org/plosone/s/data-availability

[17] For an alternate consideration based on achieving reproducibility of published results, see the Transparency and Openness Promotion (TOP) guidelines outlined in B.A. Nosek, *et al.*, Science Vol 348, Issue 6242, p.1422. DOI: 10.1126/science.aab2374

and sets forth the requirements on what associated information must also be preserved with the data to make it accessible and understandable to potential users. Figure 1 shows one characterization of different attributes a dataset can acquire, based on the sophistication of the information preserved along with it and how it is preserved. Here, the attainment of each successive level of dataset use implies that those levels below it have also been achieved, *e.g.*, in order for a dataset to be "Accessible," it must first be "Preserved."

Beginning at the bottom of the pyramid in Figure 1, at the most basic level of data curation:

- "Stored" data exist in some form, somewhere. They are not necessarily archived, nor is the storage medium guaranteed to exist in the future.
- "Preserved" data are stored in a standard archival format and placed in a location such that the bits will remain readable some time in the future.
- "Accessible" data are stored in a location in which the data can be retrieved by relevant end-users

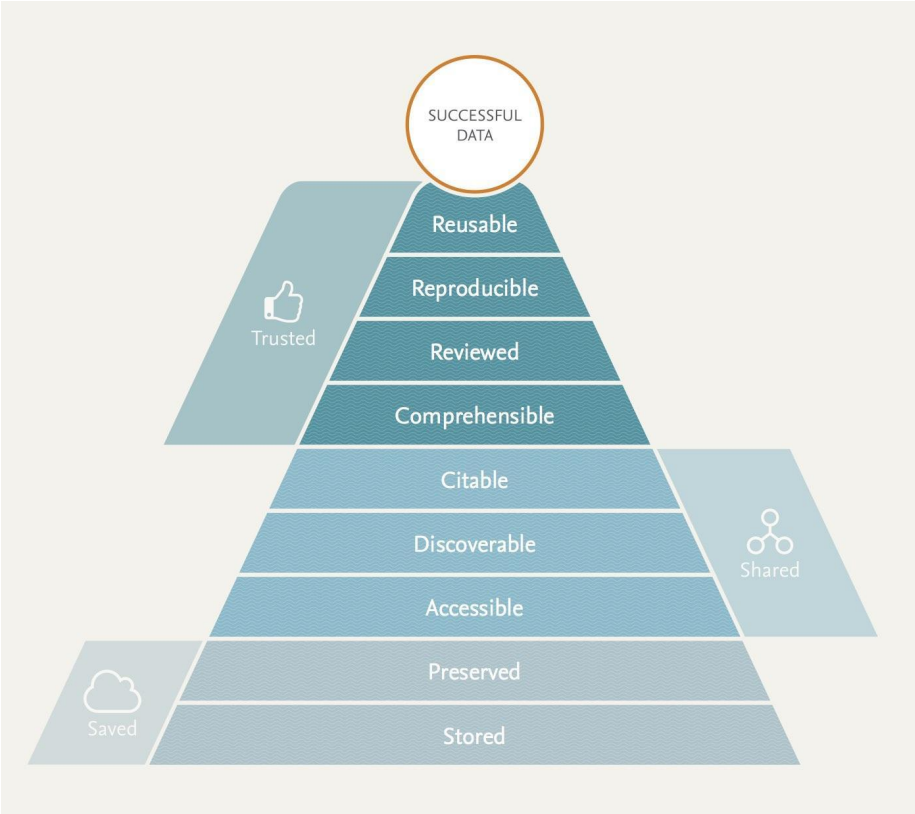


Figure 1: The attributes a dataset can acquire based on the sophistication of the information preserved along with it. Figure by Anita DeWaard.

- "Discoverable" data are indexed in an appropriate manner and with appropriate metadata enabling them to be found by searching some relevant catalogue
- "Citable" data have an identifying marker, such as a DOI, that allow them to be referenced, and allow their impact to be measured.

- "Comprehensible" data carry with them accompanying documentation of their content.
- "Reviewed" data have their content and provenance vetted by expert opinion(s).
- "Reproducible" data are provided with enough information, algorithms, software, etc., to allow the data to be re-derived.
- "Reusable" data are provided with enough information, algorithms, software, etc., that enable the data to be processed or analyzed in a different manner than that which produced the original data or results.

As these descriptions show, making data increasingly useful for potential future re-users requires investments in preserving and sharing successively more sophisticated and varied information and artifacts associated with the data. Therefore, when setting goals for the desired level of data re-use for open access data, the effort required of investigators must be a careful consideration.

A second discussion around data sharing is that of levels of data (and software), and is described by the second pyramid shown below in Figure 2. The data that finally appears in a publication have almost certainly progressed through several stages of processing, reduction, and analysis. To what extent should the results of an intermediate processing stage, and the software needed to understand it, be preserved and shared? Clearly, the answer to that question is highly dependent on the scientific result itself, the relevant scientific domain, the effort required for preservation, and on whether or not the data is intended for re-use. Sharing the numerical results backing the figures of a given publication allows comparison and potentially the combination of these results with other research results. Sharing the whole dataset from which these results were derived and the software to read it enables re-use. The level of data to be shared depends on the goals of the open data process and the level of complexity of re-use, given the auxiliary information that must be provided. Existing domain-specific use cases and practices for data re-use may also be drivers in determining what data are most usefully made available in machine-readable form. These considerations are key ingredients in forming any decision as to what information should be shared.

Various levels of processing, starting with raw data, are implied. For an example, see the data processing levels defined by NASA[18].

There is ample evidence in some fields such as astronomy, that the conscientious preservation and sharing of research data have immense benefits. For example, the data in the Hubble Space Telescope (HST) archive[19] is "science ready," calibrated by fully documented and open source code and indexed with thorough metadata describing the conditions of the observations. HST principal investigators normally

[18] Committee on NASA Astronomy Science Centers, & National Research Council, 2007, p. 12; "Data Processing Levels for EOSDIS Data Products - NASA Science," 2010

[19] https://hla.stsci.edu/

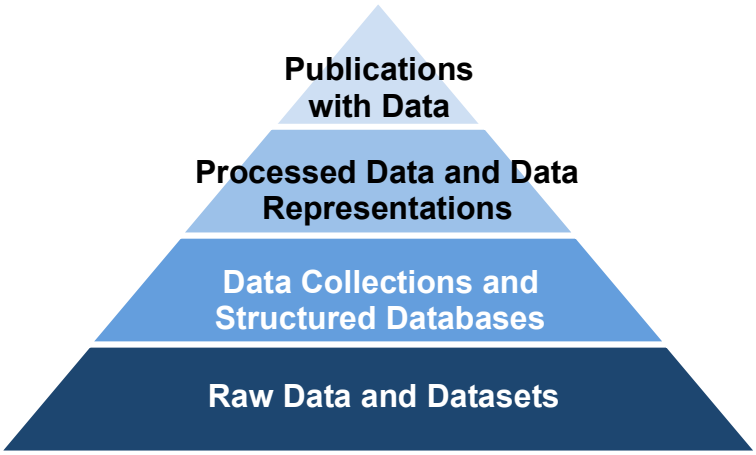


Figure 2. An illustration[20] of different data types that could support a given scientific result in a publication.

have access to their data with a one-year proprietary period, after which time the data become fully public and available for re-use. This public access has led to the majority of the peer-reviewed publications based on analyses of HST data produced *by authors not affiliated with the original proposal science team*. Moreover, the archival research papers have a similar impact in terms of citations as papers from the principal investigators.

Further evidence comes from the Sloan Digital Sky Survey (SDSS)[21], the first fully digital atlas and catalog of the sky. The SDSS was designed to present a fully calibrated, characterized, and self-consistent database of stellar and non-stellar objects. To date SDSS data have been the basis for more than 5,000 peer-reviewed papers, most of which were written by scientists having no affiliation with the Sloan collaboration. SDSS data have also been widely analyzed by citizen scientists, a number of whom have become co-authors on professional research publications.

In early 2016, a team of over 1,000 researchers used data, simulations, and processing algorithms in the discovery of gravitational waves with the Laser Interferometry Gravitational wave Observatory (LIGO)[22], resulting in a detection confidence level of over 5.1 sigma and what will likely stand as one of the most significant discoveries of the 21st century. They then shared the data, simulations, and processing algorithms (in the a number of forms, including an IPython notebook[23]) with the public to allow the analysis to be widely reproduced and understood[24].
In some disciplines such as chemistry and materials science, additional value for much

[20] Susan Reilly, *et al, "Report on Integration of Data and Publications"*, 17 October 2011, www.libereurope.eu/wp-content/uploads/ODE-ReportOnIntegrationOfDataAndPublication.pdf
[21] http://www.sdss.org/
[22] https://www.ligo.caltech.edu/
[23] https://ipython.org
[24] https://losc.ligo.org/s/events/GW150914/GW150914_tutorial.html

measured property and characterization data can be realized through aggregation of like data types. Examples of these resources include the Spectral Database for Organic Compounds managed by the National Institute of Advanced Industrial Science and Technology in Japan, or the Standard Reference Data program at National Institute of Standards and Technology (NIST) in the US. Machine-readable data can improve on common but currently manual curation practices involved in the aggregation of results for inclusion in these kinds of collections. While aggregation does not have to be a prerequisite for deposit of these data types or a direct function of general repositories, making sure that individual data sets are formatted and documented in such a way to enable automated potential and future aggregation across data types is a critical functionality for re-use. This can directly improve current community practices, refocus manual curation costs toward more value-added work and support the movement toward more open sharing of data and metadata.

Clearly, disciplinary communities have unique data management needs, and the impetus for infrastructure development must include participation and oversight from the disciplines as initiatives evolve. It may also be beneficial to future discipline-focused discussions on data sharing practices to consider more closely the different types of use cases for the open sharing of data, as suggested by the examples above.

Exemplars

We propose a number of exemplary projects and initiatives - including several NSF investments - that could be emulated to provide centralized curation resources and guidance to researchers. The following initiatives were identified as key exemplars from US federal agencies, European agencies, and disciplinary communities:

Federal Sources

- The **NIH Commons** model facilitates curation and sharing of the data products of funded research through a network of cloud providers and a system of credits distributed by NIH to support data services.
- **NASA** science centers include data archives for missions.
- The **U.S. Geological Service** publishes guidelines and resources for researchers at usgs.gov/datamanagement/.
- The **Department of Energy** is promoting computational materials design through development of open-source software.
- Finally, through the Data Management Plan (DMP) component of **National Science Foundation** proposals, the researcher is guided to foresee the management, storage, and preservation needs of the data, software, etc. that will be generated and to plan accordingly. The potential also exists to develop infrastructures to support institutional implementation of these DMPs, perhaps introducing machine-readable functionality, incorporating standard metadata or identifier components (e.g., ORCID, FundRef), and enabling public access and searching.

European Agencies

- The **EUDAT Collaborative Data Infrastructure (CDI)**[25] brings together a network of data centers to create an infrastructure for data services across Europe.
- Similarly, the **European Open Science Cloud** initiative provides recommendations and planning for European cloud-based infrastructure and policies to broadly support open science spanning disciplines.
- The **Engineering and Physical Sciences Research Council (EPSRC)** has underscored the importance of software infrastructure development through its Research Software Engineer (RSE) Fellowship program.
- The **Wellcome Trust** is prioritizing open research as an area for development.
- The **Innovative Medicines Initiative (IMI)** also promotes innovation through exploitation of research output.

From the Disciplines

A number of initiatives that were funded to support certain disciplines have subsequently scaled to other disciplines.

- **CyVerse**[26] (formerly the iPlant Collaborative) is an NSF-funded cyberinfrastructure resource that was initially developed for plant sciences but has since successfully grown to incorporate a variety of discipline-specific datasets and communities.
- Similarly, the **Dryad**[27] digital data repository is moving toward sustainability and disciplinary inclusion through a stakeholder membership model.
- The **figshare**[28] digital data repository has been embraced by a wide number of publishers of articles in the NSF MPS environment, allowing publishers[29] and institutions[30] to provide branded portal access to publication affiliated resources shared on figshare, as well as better data tools and services to authors of research articles, linking back to the version of record.
- The **HUBzero** platform[31] for collaboration accommodates construction of discipline-specific hubs for team science, including open-source tools for communication and publication of data and other research products.
- The **Open Science Framework**[32] facilitates transparent research through

[25] https://eudat.eu/eudat-collaborative-data-infrastructure-cdi
[26] http://www.cyverse.org/
[27] http://datadryad.org/
[28] https://figshare.com/
[29] https://figshare.com/services/publishers
[30] https://figshare.com/services/institutions
[31] https://hubzero.org/
[32] https://osf.io/

publication of project plans and outcomes with unique identifiers. Their partnerships with various research organizations has resulted in SHARE[33], a platform for sharing research results.

- Successful collaborations implemented by universities include **Notre Dame's CurateND[34]** repository and funding model for allocation of computing resources, along with its affiliation with the Open Science Framework.
- The **NIST Thermodynamics Research Center[35]** is an example of public/private partnerships for sustainable data management through industry relations. Several publishers work with NIST collaboratively to ensure data integrity and management. One example is the **Journal of Chemical Engineering & Data**, published by the American Chemical Society. This journal requires all authors to use **NIST's Thermolit** program[36] to identify relevant data within the NIST archive with which their new data can be compared.

Outreach to MPS Fields and Community Discussion

This report is intended to represent the consensus opinions of the broader MPS community. A draft version of this report was circulated after the first workshop in 2015. Comments were received from many different communities. Of note were two sessions at ACS meetings and a presentation at the April 2016 APS meeting that garnered attention. In particular, the leaders of all of the APS Units were asked by the APS Office of Public Affairs to comment on the draft report; their responses were shared with the workshop working group and will form part of a response report. Previous to the second workshop in this series in December 2016, the workshop organizers collaborated with the APS to administer a survey that collected data on attitudes toward open access to data and software, prevalence of data and software sharing, sizes of shared datasets, etc. A discussion of these results and a comparison with other surveys with similar focus is presented below.

Feedback from American Chemical Society (ACS Sessions)

The draft report from the first workshop was discussed at two separate ACS national meetings in 2016: two sessions on global initiatives and funder impact on research data in San Diego in March, and an open data forum in Philadelphia in August. Many points were similar to those discussed below in the context of the APS replies. Other responses, however, reflect slightly different emphases in the culture and the infrastructure needs of the chemistry community compared to that of physics. The main points are noted here.

---

[33] http://www.share-research.org/

[34] https://curate.nd.edu/

[35] http://trc.nist.gov/

[36] http://trc.nist.gov/thermolit/main/home.html - home

- Stakeholders in the chemistry community lack clarity in what and how much data are expected to be managed and publically accessible; they are concerned they will be challenged to meet this new requirement without additional resources, thus decreasing funds for research.
- An identified need for training in data science and curation procedures. This is obviously common to any of the MPS communities, yet was emphasized strongly by those in chemistry. As typical datasets grow, it was felt that students and postdocs do not necessarily have the tools they need to interact properly with archives and data management infrastructure.
- The commercial importance of properly curated large datasets and the value of coordinated aggregation of smaller datasets. Long-standing datasets related to the pharmaceutical industry or material science have commercial importance because they are highly curated and well-maintained. Translation from basic research into industrial production can be expanded or facilitated if the data is well-described with standardized metadata. The development of new standards and tools for metadata assignment will be needed in order to advance the re-use and accessibility of new datasets.
- Licensing publically-accessible data. If a dataset is public, who owns the intellectual property, and how can misappropriation be addressed?
- Concern over a potential proliferation of standards. Even if a baseline recommendation on public access is to supply the data supporting the figures and tables in a publication, this leaves each publisher to create its own guidelines for formats, data quantity, etc. Of course, requirements that vary by funding agency could pose even more severe challenges.
- Sustainability. How are the costs of long-term storage and public access borne, and by whom? Obviously, this is a central question in this whole discussion.
- Standardized metadata for specific instruments. Many reported results depend critically on instrument type and calibration. Yet, very few standards exist, nor is there typically a manner to attach this information to a dataset.
- With respect to sharing data, historically, this has been managed through inclusion of data in supporting information or by deposition to repositories. Publishers are collaborating around the development of a site to help interested authors of published articles understand how to share published content, see howcanishareit.com[37].
- Scientists in the chemistry community look often to resources such as the ACS Ethical Guidelines to Publication of Chemical Research[38] for guidance, noting Section B around an author's obligations addresses the need to present an accurate and complete account, including data, provisioning of that data and materials when requested, and submission of data to public databases.
- What is the structure of the public access landscape? There are many interested parties third-party archives, publishers, institutional archives, government-sponsored archives, funding agencies, international standards organizations,

[37] http://www.howcanishareit.com/

[38] http://pubs.acs.org/userimages/ContentEditor/1218054468605/ethics.pdf

scientific societies, etc. How do they interact, and what should be the role of each?

Feedback from the APS Office of Public Affairs & APS Leadership

After the first workshop, a draft version of this report was shared with the Office of Public Affairs of the American Physical Society (APS). In response to this initial discussion, a questionnaire was sent to all 22 of the Unit Leaders of the APS sub-fields asking them to comment on the preliminary conclusions of this report. A document from the APS describing the questionnaire responses will be released separately. We have been given permission to quote preliminary points of concern here. The following concerns arose in the responses of the APS Unit Leaders to questions about open access policies and the draft report of this working group:

- Lack of clarity in what and how much data are expected to be included under an open access mandate.
- Placing additional significant requirements on researchers to store and prepare files for open data without a concomitant increase in resources.
- The challenge of widely varying data intensity and disparate levels of effort in implementing open data, across a vast array of scientific fields and Federal agencies.
- Potential for misunderstanding data sets without significant additional context and effort.
- Potential difficulties in maintaining access to data as digital archiving methods change over time.
- A lack of accurate estimates of all the costs involved with setting up and maintaining an open data system, including the time and effort of researchers and the administrative and technology burdens on research institutions.

Substantive discussion of each of these points will be included in the APS report. Finally, the report makes several recommendations, quoted here:

> "The APS believes that the appropriate balance of costs and benefits of an open data mandate would be to require, at this time, only the data presented and referred to in a publication to be uploaded to an open data system.
>
> Such a policy would support the goals of creating openness for both the public and researchers while at the same time ensuring that the mandate does not place an undue burden on the scientific community.
>
> An open data policy must carefully balance the benefits of providing useful data to the public with the costs and infrastructure requirements. Should OSTP find a need for a more rigorous open data policy than the one recommended above, APS urges that the federal government carry out a thorough examination of the costs and impact. This should include evaluation of the effects not only on individual

researchers, but also on national labs, user facilities, and industry. Further, APS supports the proposal in the NSF workshop report to implement pilot projects to evaluate and develop the infrastructure required for an open data system."

The APS Survey

A survey containing 11 questions on open access policies and practices was submitted to all APS members who have been first authors on APS journal publications, a total of approximately 5000 people. An additional 5 questions collected information on fields of practice and funding sources for the individual's research. More than 500 responses were garnered by early December 2016. The full survey results will be published by the APS; the survey questions are given in Appendix II. Some highlights are presented here.

- Sharing is common, but not universal:
    - More than half (56%) of the respondents have made public the data for figures, tables, etc. within the last three years. A third have shared processed data (30%) or software (33%) in the last three years,
    - 80% have shared their research data or software with another group within the last three years. 73% of respondents claim that another group has shared data or software with them in the last three years.
- There are scientific benefits to sharing:
    - 34% of respondents have published a journal article based on shared data or software from another group within the last three years; 25% claim another group has published a journal article using the respondent's data or software
    - 60% of respondents already share their data/software or would make it publicly accessible if the funding to do so were available.
    - 50% agreed or strongly agreed with the statement "My research has benefited from the current practices of data and software sharing in my field."
- The resources needed for sharing data/software can be significant:
    - 24% of users have processed datasets that exceed 100 Gigabytes in size; some have substantially larger datasets.
    - 72% of respondents stated that they either were already providing access to the data in published figures or tables or could do so with existing staffing. That figure dropped to 49% in the case of processed data, and to 29% for raw data.
- The access and preservation landscape is uncertain:
    - Respondents were asked in their estimation which, if any, of the following has the infrastructure required to provide long-term public access to research data: their research group, their home institution, their funding agency, third party repositories, journal publishers. None of these choices garnered more than 50% affirmative responses.
    - Approximately equal numbers of respondents (42%) agreed or strongly

agreed with the statement "I have sufficient access to underlying data, software, and documentation in order to reproduce, reuse, and/or validate experimental or theoretical results in my field" as those who disagreed or strongly disagreed (39%).

- Only a small majority of respondents had established policies for archiving raw data (54%), processed data (55%), software (52%). The reported rates for the software environment (29%) or analysis workflows (32%) were substantially smaller.

Demographic data confirmed that the survey respondents were a reasonable representation of the overall APS community. A substantial plurality of those who answered the survey works in condensed matter physics. There is a good balance among tenured and untenured faculty and postdocs among the respondents, as well as a fairly even split between those reporting funding from the Department of Energy or NSF.

In addition to the questions with fixed answers, an "open comments" box was also part of the survey. The responses entered there by the respondents reinforced the views that emerged from the other survey results. In addition, there were several themes raised there that were not reflected in the questions that comprised the survey. A number of respondents pointed out that sub-disciplines have their own cultures around the sharing of results and software. There were concerns about maintaining intellectual property, either in the form of software, which could be the primary intellectual contribution to a given publication, or in terms of the capability to impose an embargo on public access to data in order to give the researchers who produced it the first chance to exploit it. Many worried about the excessive burden of any potential mandate for wide scale sharing of data and software, the lack of infrastructure to support such an endeavor, and the existence of a scientific benefit in so doing. Concerns were also raised about the misinterpretation of data and what responsibility researchers may have to correct it. Many respondents also expressed a strong support for public access to scientific results without reservations. Such a mix of topics and concerns has been reflected in many encounters with colleagues as well as in the discussions that have take place at the two workshops in this series.

Comparison with Other Surveys

Several other surveys addressing similar issues have been conducted in the past few years. See Appendix III for an annotated list of Additional Surveys of Interest. Of particular note are two recent surveys that were examined and compared with the MPS survey during the second workshop: The Open Data Survey run by Springer Nature in association with Figshare and Digital Science on awareness and use of open data [39]and its accompanying *State of Open Data Report*[40] were discussed along with the

[39] (NPG), Nature Publishing Group (2016): Open Data Survey. figshare. https://doi.org/10.6084/m9.figshare.4010541.v4

IEEE: Science Gateways Today and Tomorrow [41] survey. The findings of both surveys share similarities to some of the responses regarding data sharing, who bears cost of sharing, and infrastructure for sharing in the APS survey.

The *Open Data Survey* was an extensive (60+ question) study on data generation, data sharing, and research practices, and accessed a much broader scientific base, including most research areas funded by the NSF. The IEEE survey, while focused on science gateways, included several questions on data access and sharing. Taken together these two surveys paint a broad picture of those producing, reusing, and making research data more open. 74% of *State of The Data* survey respondents have made research data open at some point, and of respondents who have never done so, 90% would consider making data open in the future. This interest closely echoes the Science Gateways survey, where 75% of respondents indicated that data collections were important to their research/education work, ranking it highly alongside data analysis tools and computational tools (72% each) and their interest in being able to rapidly publish and/or find domain-specific articles and data (69%).

Just as APS respondents couldn't identify as a majority who would provide long term access to data among: their research group, their home institution, their funding agency, third party repositories, or journal publishers, *State of the Data* researchers were similarly "uncertain as to who will meet the costs of making data open". 39% didn't know, while no majority emerged among responses related to: use of own funds, general grant funds, grant funds identified specifically for purpose of making data open, institution, or funder provided solutions for making data open.

In April 2017, a survey conducted by Elsevier Publishing in conjunction with the University of Leiden[42] was released. It addressed many of the same issues considered in the APS survey, as well as those in the surveys mentioned above. It also showed consistent responses.

Additionally, a recently-released report by an NSF-sponsored Civil, Mechanical, and Manufacturing Innovation (CMMI) Workshop on data sharing[43] comes to similar conclusions.

---

[40] Treadway, Jon; Hahnel, Mark; Leonelli, Sabina; Penny, Dan; Groenewegen, David; Miyairi, Nobuko; Hayashi, Kazuhiro; O'Donnell, Daniel; Science, Digital; Hook, Daniel (2016): The State of Open Data Report. figshare.https://doi.org/10.6084/m9.figshare.4036398.v1

[41] Lawrence, K. A., Zentner, M., Wilkins-Diehr, N., Wernert, J. A., Pierce, M., Marru, S., and Michael, S. (2015) Science gateways today and tomorrow: positive perspectives of nearly 5000 members of the research community. Concurrency Computat.: Pract. Exper., 27: 4252–4268. doi: 10.1002/cpe.3526.

[42] Berghmans, Stephane; Cousijn, Helena; Deakin, Gemma; Meijer, Ingeborg; Mulligan, Adrian; Plume, Andrew; de Rijcke, Sarah; Rushforth, Alex; Tatum, Clifford; van Leeuwen, Thed; Waltman, Ludo (2017), "Open Data: the researcher perspective - survey and case studies". Mendeley Data, v1 http://dx.doi.org/10.17632/bwrnfb4bvh.1

[43] https://design.gatech.edu/cmmi-data-structure-workshop

Discussion

In response to the feedback received, both through public fora and the survey, many modifications were made to the draft version of this report to clarify or extend the discussion of various points, including the discussion of cost, intellectual property, training, exemplars, and the role of various institutions and entities in determining and supporting open access policies that were added below. The APS survey and the ACS discussions reinforced the perspectives raised during the workshops themselves, lending added confidence to our conclusions.

Summary of Feedback from MPS Researchers

The potential benefits of public access and sharing of data are recognized by many segments of the MPS research community. However, during the deliberations encompassed by this project a broad consensus emerged that *the MPS research culture, data management tools, and archives infrastructures are not ready at this time for a move to requiring public access to all MPS research data*.

One exercise of the workshops was to propose concrete steps that may be implemented in order to move the MPS community toward public access. A general consensus arose around data supporting publications. Between the MPS disciplines, there is no uniform practice in terms of which data is available to readers of a peer-reviewed article. It seems like an incremental, yet significant step to consider the following as a baseline for moving forward to public access:

Data and other digital artifacts upon which publications are based should be made publicly available in a digital, machine-readable format, and persistently linked to those relevant publications.

This rather simple procedure is already common in many disciplines. In many cases, infrastructure exists on the publications side for the deposition of supporting data that can be linked to publications (such as Scholix.org). However, elevating this to a requirement for publication brings certain advantages. It allows more detailed peer-review of supporting data and methods, bringing more confidence to the published results. It allows other scientists access to the numerical results for ease of re-use and comparison, increasing community engagement. It places the focus on the highest-priority data, as viewed through the eyes of the researcher.

A second conclusion that can be drawn from the discussions presented above is that, unsurprisingly, *different disciplines have a wide variety of current practices and expectations for appropriate levels of the sharing of data and other scientific results*. A discipline-specific policy discussion will be required to determine an appropriate level of preservation and re-use.

It was clear from the workshop discussions that any conversation about data sharing immediately raised the issue of discipline-specific norms in data sharing practices, historical context, data complexity, analysis techniques, and funding. This is also reflected in the numerous comments received in the surveys and the discussions at other fora. From a policy perspective, the complexity of the research enterprise represents the most challenging aspect of defining a uniform level of preservation.

A third conclusion is that *the provision of public access to data entails costs in infrastructure and human effort, and that some types of data may be impractical to archive, annotate, and share. Cost-benefit analyses should be conducted in order to set the level of expectations for the researcher, his or her institution, and the funding agency.*

One of the strongest messages heard from researchers was that the extra effort required to prepare research data, software, documentation, etc., for sharing comes at a cost to the overall research enterprise. The question of who pays this cost, and whether the benefits of re-use exceed the original investments are not yet resolved.

Finally, there was a consensus that *creating incentives toward the sharing of data is the primary way to accomplish broad adoption of open access to data as the norm.* Activities related to the preservation of data and other research products for public access are not sufficiently recognized and encouraged within current reward structures, nor are there sufficient tools to make data preservation and sharing easy. Incentives can come in many forms, from mandates by publishers to changes in tenure requirements to funding policy. Some combination of many different forms of inducement will be necessary to move entire research communities towards widespread data sharing.

Steps Toward Public Access

Clearly, the initial step of linking supporting data to publications would be a first step along a path toward more open access to research data and findings. Moving beyond this, however, opens up a host of questions that should be considered and resolved. Some of these lines of inquiry require research and development, and could be seen as avenues to open access if they were funded in order to develop the requisite insight and infrastructure. The overwhelming consensus of the workshop attendees and the MPS research community was that the existing infrastructure is not sufficient to enable open access to research data. To advance toward a *potential* goal of making research data reusable, new capabilities for data and knowledge archiving systems would need to be developed. Suggested components and a potential roadmap to achieve their creation form the remainder of this document.

Elements of Open Access: Infrastructure

Ingredients

In order to approach the goal of open access, it is useful to outline the different components and infrastructure that may be required. In terms of components, to reach the highest level of re-use, the following elements of a research project should also be

preserved beyond just the data:

- **Application software:** the software used to capture or create, process, and analyze the data.
- **Workflow:** instructions, frameworks, scripts, or other high-level code used to capture data and metadata and to run the application software.
- **Software environment**: a specification or instantiation of the requisite low-level software and hardware, including operating system, architecture, libraries, machine state, etc., that are necessary to run the application software and workflows.
- **Simulation capabilities:** the capability to run the application software with different parameters than used to generate the original data.
- **Documentation:** a description of the application software, workflows, standards, and other information describing how the data were created, derived, processed, and analyzed, and validated.
- **Data characterization:** documentation of the data themselves (formats, content, provenance, etc.) and the metadata that describes them and makes them discoverable and re-usable.
- **Measurement parameters:** sample metadata, instrument metadata, measurement techniques, calibration, reference standards, etc.

The depth to which these elements need to be included will likely depend on a number of factors, including the ambition of the open-access policy, the nature of the research, and existing domain practices, among others.

Technical and Infrastructure Requirements and Capabilities

The infrastructure necessary to enable open access to data will be composed of many distinct components. The data and accompanying information must be stored in archives that have the capacity, connectivity, search-ability, and potentially, computing power to support long term open access and re-use of the data. To this end, a number of technical attributes or requirements and capabilities were discussed for open-access infrastructure. The capabilities listed below were considered critical for any universal infrastructure to support open access to data.

First, at a base technical level, the infrastructure should have the following components:

- **Federated storage infrastructure:** data will be stored in a variety of archives; these archives should be globally accessible
- **Links between publications and research data/software**, providing a means of tracing the origin of published results
- **Means for revision/correction and versioning** of archived material
- **Support for general and domain-specific data standards** as they exist
- **Open/uniform formats for instrument data**, in order to allow sharing,

interpretation, and interoperability of raw data from a variety of common scientific instruments and persistent identifiers for each instrument/sensor

Advancing in terms of complexity, the following attributes are also desirable:

- **Infrastructure for software/environment preservation** linked to the datasets
  - In order that one can determine what software and system is needed to access a data collection of interest
- **Data quality assurance infrastructure**, which would insure that the data deposited, as well as, potentially, the accompanying software, are properly saved in a structured and readable manner in the repository. This would require some
  - *automatic validation of data and results*, based on some algorithms proposed by the researchers themselves or developed by domain communities
- **Global search capabilities:** the archive, or more likely, coalition of archives has a quasi-universal metadata description that allows a researcher or citizen-scientist to search for instances of data of interest. This requires support for the researcher, for example, in terms of
  - *automatic metadata generation*, which could guarantee uniform metadata from archived datasets, as well as the capability to generate metadata automatically during the research process

  and, for the archive itself,
  - *appropriate discovery tools*, that allow scientists, industry, and the public to explore the curated data and understand its meaning, structure, provenance, and applicability
- **Machine-accessible retrieval options** that allow programmable access to data and metadata for further re-use, such as for data analysis or aggregation – via APIs, parsable URI schema, machine-readable metadata stacks, etc.

Elements of Open Access: Normative and Policy Considerations

Merely developing infrastructure for open access, however that might happen, does not guarantee that the research community will embrace open access to data as the norm of scientific behavior. Any imposition of an open access policy, for example, should be accompanied by a cost-benefit analysis in order to set the level of expectation for the researcher. This will be discipline-dependent as the benefits resulting from open access to research data are difficult to quantify broadly. This is likely due in part to the minimal quantity of data that has been open for public access in some fields and the lack of clearly defined use cases to initiate practice. In addition, a general cost model for calculating the benefits of broader access to scientific results is ill-defined.

A point of general agreement, however, is that if the process of doing science in the era of open access becomes *easier* because of the tools introduced for data/knowledge

preservation, then widespread appreciation and adoption would quickly become the norm. In other words, if there were an "economic" incentive that made the process of doing science more productive and, as a by-product, made the preparation and release of research data and software easier, there would be little resistance to this change in focus. This may be an impossible goal. However, with current trends toward complex data analysis workflows, there is an increasing call for tools that allow the preservation and reproducibility of an analysis merely so an individual scientist can remember and restore what he was doing the previous week. The extension of such tools to a common architecture or architectures that allow preservation and sharing based on common standards would be of great benefit. Investigations in this area may provide a fruitful space for advancing the goals of open access.

Incentive Structure

Beyond the success that might be engendered by widespread adoption of tools that make knowledge preservation and archiving easier, other incentives will also be necessary. Perhaps a way forward is to focus first on data preservation and archiving in a "useful" manner. If the problems associated with storing large quantities of data in a federated archive in a manner that renders it globally discoverable and searchable could be solved, the additional complexities and knowledge required for re-use could be layered on top of the data infrastructure. In this way, the usefulness of the stored data would increase over time as more information and more tools for re-use are added.

Incentives must be established to encourage researchers to engage in these activities beyond just doing the minimum necessary to satisfy a requirement. For example, small amounts of additional funding or some kind of data preservation "credit" could be awarded for datasets stored in a re-useable way. Changes to the reward structure for publishing such that data creation and data publishing citations are recognized as a valuable contribution to the scientific process would support these endeavors. In addition, removing *disincentives* for publishing data is also important. For example, providing a means to easily embargo public access to a dataset while sharing it internally with collaborators might make use of a central archive more attractive.

Providing additional incentives is obviously relevant for the broader and more complicated problem of knowledge preservation. It is interesting to note, however, that some funders and publishers have had success in moving communities towards data sharing and open access to results by imposing mandates for these practices. Since the number of examples for this is small (but growing), it is not clear that communities in question had already established widespread practices that would enable them to respond positively to such mandates, or whether the mandates were responsible for pushing the community in the direction of more openness.

Establishing Norms of Community Behavior and Policy Guidelines

If open access is to become the standard, a host of issues surrounding the curation of data, outreach, training, workforce development, etc., need to be addressed. Some of

those that should be considered are listed below. Here, we use the generic term "data" to refer to all aspects of a preserved and shared research project.

- **Establishment of best practices in data management**: how is the data stored? What is the required time frame over which stored software, virtual machine images, etc. will remain executable? What is the expected storage lifetime?[44]
    - Through what review process are these criteria established?
- **Establishment of ways to quantify the usefulness of data**: if a reward structure is to be established for data creation and publication, metrics are needed. These could include usage and impact measures, such as downloads, accesses, citations, etc., as well as peer review or other forms of validation and evaluation.
- **Establishment of a culture of data citation**: this may already exist, but is likely to become much more prevalent as the amount of openly accessible data increases.
- **Establishment of a de-accession policy**: when can a dataset be declared "obsolete"? Who decides?
- **Establishment of a policy for preserving data for non-published experiments:** should *all* data be published? What are the limits of accessibility? How can one characterize a dataset as "scientifically-useful"?
- **Establishment of a communication structure for published data:** how should other researchers be notified of the publication of a new dataset? It's possible that a service that provided some sort of periodic catalogue of new and interesting datasets might dramatically increase reuse.
- **Establishment of training/workforce development programs:** materials that introduce and explain the available tools and the analysis structures and present preservation best practice will be necessary for students, postdocs, and other professionals to understand how data should be preserved and shared.
- **Establishment of licensing practices for data, code, and workflows:** standards that preserve scientific norms of verifiability, re-use, and citation.

All of these are very broad issues. It is likely that the policies and structures that are eventually established and accepted will grow up organically over time, since all of them are, at base, issues of evolving community practice. We list them here as potential ingredients for a successful open-access culture. Clearly, targeted infrastructure investment in a manner complementary to other activities in this area would help in hastening the establishment process and for community building.

Licensing, Copyright, and Intellectual Property

Digital scholarly objects fall under intellectual property law. By default, original works of authorship in fixed form, such as software, figures or tables, and possibly data, are

[44] Note that some of these questions are related to those around what constitutes a "trusted repository." This is a slightly higher-level discussion than that functionality, however.

subject to copyright. Copyright is a legal barrier to reproducing the work *without permission*, for example copying it to another computer, and to creating derivative works, such as modifying existing software code for use in a new research setting. Additionally, such objects may be patentable, e.g., some software. Providing a default open licensing structure, or set of recommendations, for work funded via the NSF can facilitate the legal re-use of these digital scholarly objects including the verification of published computational findings. We suggest using some guidelines such the Reproducible Research Standard[45] as a starting point for this discussion.

A related question raised repeatedly was the implications for the researcher's rights to intellectual property if all research results are required to be public. In cases where a substantial portion of the effort on a research project goes into producing software or other products with high future value to the researcher, there is naturally a reluctance to immediately make all methods publicly available. This is an important issue, but one on which no concrete conclusions were reached during the workshops.

Elements of Open Access: Training Aspects of Changing the Research Paradigm

While the open access movement has been supported and enabled by many in the scientific community, still more are holding on to the traditional research and publishing paradigm. The 'publish or perish' model has a common modality – introspective research – a competitive environment where all research data/findings are treated as intellectual property and dissemination of research is very tightly managed. It is this environment that trains future researchers to be closed to sharing their data and discourages community evaluation of data/results – to the potential disadvantage of the researcher and the community.

A move to open access requires retraining of researchers, not only relative to the perspective above but also relative to how they regard and treat the digital assets they generate during the research process. This retraining will need to include the fundamental realignment with open access/data core requirements and integration of these practices with the scientific process. Because this realignment is coupled with the inherent digital nature of science and scientific data, a second set of skills based around data science are also needed by researchers. Below, we outline a set of questions that touch on the important ingredients of a training program that would prepare researchers for a scientific process based on open access to results. A related discussion prepared by the National Academies[46] reaches similar conclusions.

- What expertise and skillsets are needed?
    - An understanding of
        - unique identifiers for various elements in the process:

[45] See http://scitation.aip.org/content/aip/journal/cise/11/1/10.1109/MCSE.2009.19

[46] https://www.nap.edu/download/18590

        - researchers/authors, publications, institutions, instruments, chemicals, etc.
        - preservation/archiving of knowledge
        - file formats (open vs. proprietary)
        - data and software tools & services
        - appropriate metadata to describe critical scientific and technical aspects of the data
    - An appreciation of
        - open/agile software development
        - data science (informatics/computing/statistics)
        - digital curation (including validation)
        - data standards
        - the scale of data
        - reproducibility issues
        - analysis requirements
    - A consideration of the importance of ethics related to data
        - manipulation of data
        - anonymization of data (e.g., implementation of FERPA requirements)
    - An understanding of the intellectual/legal issues, copyright
        - copyright of data/datasets
        - licensing agreements from publishers
        - data embargos
    - Working in a team environment/collaboration

Clearly, the depth of training in these issues varies with the level of specialization of the researcher. Data scientists working closely with archives will need a more complete skill set than a graduate student or an undergraduate. All researchers who produce research data and scientific results, however, should have an exposure to and basic knowledge in the areas outlined above.

- How do we train researchers?
    - Formal education (in a standard/updated curriculum and/or incorporated into discipline curriculum)
    - Informal training (continuing education, industry)
        - e.g. Software Carpentry, Data Carpentry, Lab Carpentry
    - "On-the-job" training
        - learning through researching
        - internships, Fellowships, Residencies
        - collaborative development

A growing set of resources exists for training in the areas outlined in the first list, above. Educational programs in Data Science are proliferating, for example. Outside of a fully immersive experience like a dedicated degree program, however, there are few opportunities to obtain a basic introduction to the necessary material that would

complement discipline-specific training in scientific research. Most training at the interface of science and data science occurs in "on-the-job" settings.

- Who do we train?
    - Faculty
    - Academic research staff
    - Industry & National Lab research staff
    - Postdocs
    - Graduate students
    - Undergrads
    - K-12 students
    - Librarians (information professionals)
    - Specialized staff, e.g.
        - research software engineers
        - data scientists

For a reorientation of the research process, all levels of researcher, including those who will feed the personnel pipeline, will need exposure to the training and techniques that will support a future open science paradigm.

It is clear that such a large paradigm shift will not happen overnight and many questions remain. Who is responsible for training, promoting these skills, developing training materials? How do we support single PI projects, where the PI has to be somewhat of an expert on all (especially at undergraduate research institutions)?

We must also emphasize how to integrate data related activities into the scientific process, make researchers aware of concepts relative to specific parts of the process, and encourage collaboration/team-building as it is impractical to suggest that all researchers will have all data related skills.

The Open Access Landscape

In the context of the broader discussion of open access infrastructure, it is useful to outline potential actors in this domain and what their various roles could be in creating and sustaining that infrastructure. While some roles may be obvious, considering potential contributions may help guide further discussions on what kind of open access infrastructure might eventually evolve, and how the various actors might be able to catalyze development. In particular, we might define the potentially unique roles that NSF can play in this enterprise. As we discuss each actor, we will attempt to catalogue the roles that could be played, the opportunities each might have to promote change or development of an open access ecosystem, impacts on time and costs, and possible mitigation strategies. There may be many potential opportunities to partner across multiple stakeholders to improve the culture and infrastructure for depositing research data.

The Researcher

Roles: Clearly, the researcher is responsible for generating the data and the associated scientific results that are at the core of this discussion. This activity would then include data collection and analysis, publication preparation, and the preparation of data, software, and documentation for deposit in an archive. This last set of activities would then include the generation of appropriate metadata to describe the results and the process of producing them. If the research is federally funded, the researcher is responsible for generating and adhering to a Data Management Plan. The question of publication raises a number of issues, including intellectual property and choice of journal. Often, a researcher is responsible for training her students and postdocs not only scientifically, but also for proper data management. The researcher is also partly responsible for managing the lifecycle of her data: when, for example, should a dataset be de-accessioned?

Possible avenues for catalysis: With proper tools for the preparation of data, software, and documentation for archiving, the researcher could be the solution to many of the vexing problems associated with open access. If the researcher were provided with a "frictionless" environment that both enabled easy data sharing and maintained (or increased) research productivity, many of the obstacles to open access would be removed. Without the researcher's cooperation, the goal of open access will be difficult to achieve. Researchers also serve as role models for each other and their trainees. Establishing and propagating best practices is a local process.

Cost Drivers and Mitigation Strategies: The researcher, as the source of the data, will be required to format and document with metadata all research products. An APS survey from November of 2016, "Data Practices in Physics", provides anecdotal evidence of about 120 hours required to prepare 200 GB for release in one case, and 1 month for 1TB of data in another. The required effort will decrease as this becomes a norm in the field. For example, tools will be developed by data archives to improve the data ingestion process, and instrument makers may improve their tools and workflow to automatically generate metadata. Lowering the barriers for individual scientists to get their data out there is huge factor in mitigating reticence to share, but getting data out in a form that is truly re-usable in a machine driven world presents challenges and often costs for developer, instrument, archive and vendor communities. It would be helpful if the same archives and tools could be used to archive data not associated with a paper. The level of detail of metadata required for archiving should match available tools. There will be minimal costs associated with proposals: templates for standard procedures in each field would be helpful to minimize this. Researchers may comply with funder and institutional open data sharing policies by sharing data on their grant funded or institutionally funded personal websites, through institutional archives, or through deposit to disciplinary or agency archives. In the first case, grant funding cycles and careers that start at one institution and end at another with stops in between make the personal website a poor choice for data sharing. Deposit to agency or funder provided repositories may not be available to all research projects and researchers. Therefore, organizations and archives like the Center for Open Science,

FigShare, Dryad, and disciplinary archives occupy a popular middle ground that mitigates the disruptions of grant funding cycles and institutional affiliation. Funding for such repositories plays an important role for researchers.

The Archive

Roles: As a long-term repository of scientific knowledge, a repository is responsible for the appropriate ingestion, storage, and curation of a dataset and its accompanying documentation. Typically, an archive provides access to stored data by providing a means through which the data can be discovered, such as a catalogue. Archives control data access through a set of access policies, and insure the integrity of the data through security controls. Archives may provide persistent identifiers for data that they ingest, and may provide persistent links between data that they hold and other external documents or information. In terms of internal data control, services such as embargoing ingested data or data version may also be provided. Archives are also partially responsible for data life cycle management.

Possible avenues for catalysis: Since the archive often serves as the initial entry point for the researcher into the preservation domain, archives can develop pieces of the "frictionless" infrastructure, such as applications facilitating data ingestion, metadata services, etc. Domain-specific repositories can play a particularly large role here, since they could work closely with researchers to develop ontologies and metadata schema tailored to the researcher's needs. Depending on the provider and the target population, archives have different stakeholders, which can lead to a variety of possible partnerships to develop this infrastructure.

Cost Drivers and Mitigation Strategies: Many research data repositories are built on open source platforms that demand significant human resource to stand-up, maintain, and keep current. University funding for such archives could ostensibly come from grant overhead but in practice support for such archives is funded unevenly at best. Government Agency, non-profit, funder, publisher/vendor, disciplinary and society provided archives each have their own cost drivers, ROI and commensurate variance in quality, feature-set, and funding stability. In particular, cost models for archives need to take into consideration the definition of "long term" data preservation: whether policies dictate preservation for five, ten, or fifty years drives not only the cost of storage, but also the cost of format migration, platform upgrades, and maintenance. The multiplicity of repositories among stakeholders results in duplicative infrastructure costs. There is potential for cost savings through disciplinary, consortial, and funder-provided shared repository infrastructures. Efforts to provide federated search and interoperability between repositories, and funding to support such efforts help mitigate the difficulty of searching across the often-siloed landscape of archives. Several well-developed cost models exist for identifying archives' cost drivers. The 4C project carried out a review of ten existing curation and preservation cost models and developed a cost model comparison table[47] and a web site that provides access to the

[47] 4C http://www.dcc.ac.uk/projects/4c/cost-model-comparison-table

models[48]. These are valuable tools for those budgeting for or those evaluating the budgets of particular archives. Another useful tool for identifying the cost drivers for archives is the APARSEN *Report on Cost Parameters for Digital Repositories*[49].

The Researcher's Institution

Role: The institution within which a researcher resides should serve as an advocate for the scientist and provide support and infrastructure. In particular, institutions could be repository providers, supporting the function of long-term preservation. Institutions often provide training or other professional formation. Institutions, being directly responsible for the researcher, are most often the standard bearers for research ethics. For federally-funded research, the institution is a partner with the funding agency, following and administering grant agreements. Many institutions maintain consultants who advise them on how they may affect federal policy.

Possible avenues for catalysis: Institutions, especially universities, establish and maintain a reward structure that determines the emphasis of their faculty or employees. Currently, quality of publications, number of citations, and (potential) grant income receive the most weight in considerations for hiring and promotion. Institutions also have the capability to change the reward structure. If more weight were to be placed on the production and citation of datasets, for example, then their creation and curation would immediately become more prevalent. Changing the reward structure is a very large lever that one could use to encourage practices related to open access. A second area in which institutions could catalyze change is innovation in local infrastructure. Institutions tend to be extremely conservative where IT infrastructure is concerned. Supporting innovation in this area, especially related to archives and their operation and interoperability could enable the rapid development and adoption of new strategies for long-term archiving and curation.

Cost Drivers and Mitigation Strategies: University provided research data repositories aka digital archives are meant to provide accessible, long term archiving of scientific research objects that are discoverable and ready for citation and scientific re-use, yet there are few safeguards to ensure that all have sufficient and/or stable, dependable funding. Funding for such efforts varies between institutions. Some universities have developed their own repositories on open source platforms over time with in-house personnel, while others' elect to use a branded hosted service like figshare for institutions, Digital Commons or hosted DSpace. Unfortunately, there is little to prevent even initially well-funded repository efforts from falling into disrepair due to lack of maintenance funding, though trusted digital repository certification processes seek to mitigate the way fluctuating funding can jeopardize a repository's preservation or access missions. Uneven funding landscapes in part account for why scientists do not all have the same experience when they deposit to institutional repositories. Some university and disciplinary repositories are feature rich, intuitive, and well-staffed,

[48] 4C http://www.4cproject.eu/summary-of-cost-models

[49] APARSEN-REP-D32_1-01-1_0

while others may have little support, aging interfaces, and unhealthy ratios of support staff to faculty. Consequently, as described in the survey section of this report, while the percentage of faculty who are sharing data or plan to do so has reached a majority, the number who report unaided self-deposit to university repositories is likely still a minority[50], perhaps because when deposit is required in publisher, funder, and/or government agency systems, researchers are less likely to concurrently deposit at their home institutions? A more sustainable funding model for provision, maintenance, and improvement of institutional data repositories would be to recommend or require that a certain percentage of overhead to be allocated to repository infrastructure.

Publishers

Role: Clearly, publishers are responsible for communicating research results to the public. In many cases, this communication already includes data reporting, either in the article or the supporting information. The peer review processes that publishers facilitate and manage bring a more uniform level of quality to the results that are published within any specific publisher or journal. Publishers establish norms of presentation style and content, and are often collaborative community leaders in developing quality and data submission requirements which can vary by publisher and by community They curate the results of the scientific process to produce, preserve, and disseminate the copy of record, including some data. Publishers curate research data in some cases, and in some cases, provide an archive in which data related to publications can be stored and preserved through dark archive initiatives such as Portico[51] and CLOCKKS[52]. They already (in large part) provide persistent identifiers for published results, and persistent links to other supporting information - primarily through the publishing industry-supported CrossRef initiative[53]. The metadata, tagging, and curation provided by publishers to search and discovery tools allows the discoverability of articles and supporting information, data. Publishers provide metrics of academic quality, such as the number of citations for a given journal article, and, in many cases, alternative metrics around usage and public impact. Kent Anderson's *96 Things Publishers Do[54]* provides a thorough listing of ways publishers add value.

Possible avenues for catalysis: Editorial policy, such as requiring that data be deposited in an archive before or concurrent with article publication, could have a dramatic effect in increasing adoption of public data access policies and norms. Enforcing the adoption of recognized standards in terms of data formatting, content, etc., can have a similar effect. In collaboration with the NSF and other funding agencies, publishers can

---

[50] SPEC Kit 334: Research Data Management Services by Fearon, D; Gunia, B; Lake, S; Pralle, BE; Sallans,A. (July 2013)

[51] http://www.portico.org/digital-preservation/

[52] https://www.clockss.org/clockss/Home

[53] https://www.crossref.org/

[54] Anderson, Kent, "96 Things Publishers Do (2016 Edition)," *The Scholarly Kitchen*, 01-Feb-2016. https://scholarlykitchen.sspnet.org/2016/02/01/guest-post-kent-anderson-updated-96-things-publishers-do-2016-edition/

play a valuable role in identifying data needs and new opportunities for projects that could be advanced together. CHORUS[55] is a model example of how publishers have already developed an innovative and cost-effective infrastructure around advancing the public access to the results of funded research. The CHORUS service is provided free of charge to federal agencies and enables readers searching the NSF Public Access Repository (NSF-PAR)[56], hosted by DOE, to follow links that point to articles in context of the journal where they were published. CHORUS supports NSF's partnership with the U.S. Department of Energy (DOE) Office of Scientific and Technical Information to provide distributed repository and search services and both use the interoperable CHORUS framework, along with CrossRef's Open Funder Registry[57], together to provide an article submission workflow for NSF grantees and facilitate public access to articles that result from NSF funded research. The adoption or support of publishing policies, such as Registered Reports[58] (pre-determining and accepting the experimental method before the experiments are conducted, or guaranteeing the publication of null results from registered studies), could impact reproducibility.

Cost Drivers and Mitigation Strategies: Publishers must maintain links between DOI's for the paper and the data. Just as they are now responsible for tracking citations for papers, they will be responsible for tracking citations to data. Maintaining discoverability and links between papers, data, and software creates an interoperability burden that is unequally and uneasily shared by agencies, publishers, funders and nonprofit organizations. Uniform policy and ways to achieve stable interoperable systems would make it easier and cheaper for all in the field (e.g. see efforts in the biomed field[59]). Publishers will also have to coordinate with archives for the peer-review process to assure the data and metadata is up to publication standards. As archives mature they should be encouraged to implement and adopt standard API's that reduce costs for publishers, authors, and users. A publisher may also take on the responsibility of providing an archive, in which case all of the costs and mitigation strategies from there will apply.

Scientific Societies

Role: Scientific societies serve as a point of contact for the broadest segment of a scientific field. Often, they support publications and repositories, so they acquire these roles as well. Scientific societies serve as the means by which the scientific culture for a given discipline is sustained. Among their activities are the development of community related ethics standards and best practices, the development and promulgation of community standards of excellence, community engagement, outreach, the identification of community needs and providing resolutions to meet needs, professional development, and advocacy.

---

[55] https://www.chorusaccess.org/

[56] http://par.nsf.gov/

[57] http://www.crossref.org/fundref/index.html

[58] https://cos.io/rr/

[59] https://doi.org/10.1101/100784

Possible avenues for catalysis: Because their reach is so wide, scientific societies can play an extremely important role in the communication of new concepts, requirements, tools, and developments around public access to scientific results. Societal prizes are an important part of the reward structure and could be leveraged for changing perceptions of worth for various data-related activities. To the extent that societies also validate curricula and support professional development, they can add elements of training relevant to data science to their programs.

Cost Drivers and Mitigation Strategies: In many fields, societies (along with funding agencies) lead the way in establishing best practices for data preservation and sharing. Initially this may include sponsoring workshops and special tracks at society meetings, and working with publishers. Societies may also convene standing committees or working groups to develop disciplinary standards for preservation file formats, metadata schemas, interoperability tools, training materials, or even repositories for their members' use and benefit. In these meetings and discussions, disciplinary e-norms can be established, shared, and periodically reviewed to ensure they are aligned with members' interests, capabilities, and expectations. Funding to societies that support such activities mitigates the gaps between what is available to funded and unfunded researchers, as well as for those at have and have-not institutions. They can also potentially provide regular, recurring, calendared opportunities for improving standards, practices and awareness. These efforts can help mitigate the irregular nature of research funding and the varying budgets and capacities of researchers at different phases of their careers.

National and International Entities

This category encompasses various organizations, such as those that set international standards policies for data access, data types, data citation, etc. (e.g., ORCID[60], CoDATA[61], FORCE11[62], DataCite[63], ISO[64], DOI providers[65]), those that set or coordinate science policy (e.g., IUPAP[66], IUPAC[67], ICSU[68], OSTP[69], NAS[70], CENDI[71], BRDI[72]), and those composed of researchers organized around common goals (e.g. CNI[73] or RDA[74])

---

[60] https://orcid.org/
[61] http://www.codata.org/
[62] https://www.force11.org/
[63] https://www.datacite.org/
[64] http://www.iso.org/
[65] http://ezid.cdlib.org/
[66] http://iupap.org/
[67] https://iupac.org/
[68] http://www.icsu.org/
[69] https://www.whitehouse.gov/ostp
[70] http://www.nasonline.org/
[71] https://cendi.gov/
[72] http://sites.nationalacademies.org/PGA/brdi/index.htm
[73] https://www.cni.org/

or geopolitical organizational or regulatory bodies. One might also consider industrial partners as actors in this category, who may be significant consumers and generators of data, and/or set community expectations for quality of data.

Role: Some organizations of this type have the capability to create and enforce new standards around issues related to data handling and sharing. All have the potential to strongly influence trends within scientific communities by creating new regulations or generating new policy. While these organizations are often interdependent, many can act independently to promote change.

Possible avenues for catalysis: Because these organizations are the standard-bearers, they can be extremely powerful in terms of generating change. Clearly, those in government also recommend or set policies, which can have an immediate effect on the direction of a scientific field, or, in this case, how scientists behave. Those connected to or governed by industrial entities can also have significant leverage toward change given the potential economic implications. A coordinated effort between those involved in developing policies and infrastructure around open access would be beneficial.

Cost Drivers and Mitigation Strategies: We have earlier identified that the plurality of repositories can be confusing to scientists, it is also frustrating for the National and International entities that sometimes fund science and /or scientific collaboration across borders. "The European Union has denounced the costs associated with funding the current plurality of online databases in biomedicine as unsustainable in the long term, and is pushing for the centralization of facilities for data sharing as a possible solution.[75]" National and International entities can work together and serve as collaborations and funding umbrellas for education and pilots related to interoperability, federated search, and joint preservation of consortially funded research objects and cyberinfrastructure. National and international entities have successfully worked together through the Alliance for Permanent Access to the Records of Science in Europe Network(APARSEN)'s *Report on Testing of Cost Models and Further Analysis of Cost Parameters*[76], which presents the results of the analysis of cost parameters, the testing of cost models, the relationship between costs and benefits, and the links to an EU Coordination Action. Consideration of how NSF's MPS funded research could benefit from similar attention to cost modelling could take the form of funding an extension of such models as well as an update and comparison that considers disciplinary use by MPS.

---

[74] https://www.rd-alliance.org/

[75] Treadway, Jon; Hahnel, Mark; Leonelli, Sabina; Penny, Dan; Groenewegen, David; Miyairi, Nobuko; Hayashi, Kazuhiro; O'Donnell, Daniel; Science, Digital; Hook, Daniel (2016): The State of Open Data Report. figshare. https://doi.org/10.6084/m9.figshare.4036398.v1

[76] APARSEN-REP-D32_2-01-1_0

National and International Laboratory facilities and instrument data centers
Role: Labs and instrument data centers are primarily funded for active data observation, collection and analysis. The capabilities of their information systems and staff can have positive impacts on data preservation and sharing but may be constrained by inefficient or burdensome infrastructure. In some cases the requirements for creating transfer packages or harvest-ready mechanisms to enable downstream sharing and preservation at various phases of a research project may introduce unnecessary complexity into the research workflow. Labs and instrument data centers create best practices and are informed by standards for data sharing within their disciplines. Methods and practices for data handling during active data collection often influence preservation and sharing decisions and capabilities further in the research lifecycle.

Possible avenues for catalysis: Because the staff and equipment at a facility is used by a large collection of researchers, anything that they provide is much more likely to be adopted by researchers using the facilities. Thus investment up front can have a multiplicative effect on the field. This must be balanced with working with the field to adopt procedures that work for everyone, not just those using the facility.

Cost Drivers and Mitigation Strategies: Development of new tools is both cheaper and most expensive at a facility. The tools must be more general so as to be useful (and easy to use) for everyone using the facility. On the other hand, the incremental work is small compared to the cost of many groups creating custom versions of essentially the same tool. Large up-front costs can be reduced by targeting smaller subsets and groups, perhaps by focusing on the most popular use cases.

Instrument Makers and Software Vendors
Role: A great deal of scientific data is collected with commercial instruments and analyzed with vended software; the data originates with these devices and thus the preservation process should also begin with them. As data preservation takes hold and metadata standards develop, instrument makers and software vendors increasingly play key roles in incorporating these standards and reducing the workload on individual researchers through open formats.

Possible avenues for catalysis: This is a chicken and egg problem - as soon as there is a demand for capabilities to aid the researcher, vendors will respond, and if the features come with little additional work researchers will adopt them. On the other hand, if data standards do not exist or are not enforced, different tools may not produce interoperable, consistently machine readable representation of data.

Cost Drivers and Mitigation Strategies: Though the instrument maker or software vendor bears the cost of development, they pass along the cost to the researcher. However this tradeoff can be mitigated by convenience and sharing features that benefit the researcher. For example, users of Mathworks' Matlab software benefit from

the MATLAB File Exchange[77] which has ~26,000 code repositories and GitHub has ~40,000 written in MATLAB code. Well-understood practices for modularity, standards for metadata and any required API's for the instrument maker and software vendor to program against help reduce costs of sharing for researchers, and potentially time to market for instrument makers and software vendors. Open, well-understood standards also potentially foster some competition and/or innovation between vendors as they seek to make their solutions more fully featured.

Funders

Role: Public and Private funders are among the primary providers the financial support for research. However, their roles are much broader than just supplying funding. In response to scientific needs, funders can provide the seeds for Research and Development to open up entire new fields of science or to develop prototypes for new infrastructure. Funding agencies support workforce development, they foster public and private partnerships, and, through overhead contributions, they support underlying research infrastructures. They communicate and enforce best practices through the proposal review process, hosting workshops, and largely serving as an entity to bridge geographically and disciplinarily separated efforts. Through requirement and review of Data Management Plans and the articulation, incentives, and enforcement of related open access and data sharing policies, funding agencies can and do communicate an ethos of open access toward data reuse that permeates into disciplinary norms. Ashley Farley, Associate Officer of Knowledge and Research Services on the Open Access Team at the Bill and Melinda Gates Foundation said in a recent interview with *ScienceOpen* that "A researcher can take lead, but will need the support of their funder, institution, and community. Stronger together!"[78]

Possible avenues for catalysis: Since funding agencies provide financial resources based on clearly-defined requirements, funder compliance expectations can motivate fairly rapid changes in researcher behavior. This is perhaps, along with the reward structure, the largest lever for inducing change in attitudes and adoption of open access and encouraging data sharing. Indeed, since the NSF began requiring a data management plan for proposals submitted or due on or after January 18, 2011, not only has researcher behavior changed, but services and tools to support that behavior have proliferated as well. Data management services are now offered at many research university libraries. Data management planning tools like the DMPTool as well as NSF specific templates help ease researcher burdens and encourage best practices. The NSF released its Public Access Plan in 2015 and now requires public access to articles and juried conference papers resulting from awards issued in response to proposals submitted or due on or after January 25, 2016**.** Similarly, the Bill & Melinda Gates Foundation also introduced a new open access and data sharing policy in 2015,

---

77 https://www.mathworks.com/matlabcentral/fileexchange/?requestedDomain=www.mathworks.com

78 Tennant, Jon, "Ashley Farley of the Gates Foundation: "Knowledge should be a public good" *Open Science Interviews*. Jan 31, 2017. https://tinyurl.com/ja7f7ch

allowing for a two-year transition period during which grant recipients could embargo their work for 12 months. From Jan. 1 2017 onward, researchers who receive funding from the foundation "must make their research and underlying data available, for example by publishing it in an open-access journal or depositing it in a public repository." [79] Meanwhile, NSF- funded researchers still benefit from an embargo period, but must "make their work covered by the agency's policy "available for download, reading, and analysis free of charge no later than 12 months after initial publication." The material must "Possess a minimum set of machine-readable metadata elements in a metadata record to be made available free of charge upon initial publication (Section 7.3.1) ", "be managed to ensure long-term preservation (Section 7.7);" and "be reported in annual and final reports during the period of the award with a unique persistent identifier that provides links to the full text of the publication as well as other metadata elements." These requirements set the stage for greater interoperability between scholarly publication and research data information systems, and open the door for greater discoverability and preservation of research outputs.

Additional activities that funding agencies like the NSF can use to promote change include communicating best practices and success stories, providing information and sharing compliance supplements or targeted funding to seed or pilot open access, data citation, data sharing, preservation and interoperability projects, and through outreach and education opportunities like early career fellowships, workshops, and center of excellence programs that emphasize the importance of open access to data. Some concrete examples are presented later in this document.

Cost Drivers and Mitigation Strategies: Clearly, a single funding agent cannot create and sustain a universal open-access infrastructure for all scientific results. The scale required is beyond the funding reach of essentially any individual funder. Coordination and cooperation between all interested parties are likely to be the only ways that this effort can succeed. Funders have the additional leverage of incentives, either through the review or award process, by which researchers can be directed towards best practices. If community awareness, capability, and consensus in a field can be built around tools and practices it will substantially reduce costs in the long run. It will also make for a field much more willing to work with a funder to implement the tools that enable best practices.

Costs and Cost Models

In presentations and discussions at the workshops, we researched and considered the applicability of various data storage calculators and data preservation cost-models (see Appendix I: Cost Models) and their applicability to MPS disciplines and their

[79] Straumsheim, Carl. "Openness by Default." *Inside Higher Ed*, January 16, 2017. https://www.insidehighered.com/news/2017/01/16/gates-foundation-open-access-policy-goes-effect-joining-others.

needs. Such calculators and cost models can be valuable tools for informing planning and policy discussions. However, for the MPS use cases we discussed in the workshops, the existing cost-models were often too simplistic or too specific to scenarios outside MPS to be useful "out -of-the-box". For example, as mass storage options in the cloud become more ubiquitous, the cost to simply store data can be estimated at as low as $0.004 per GB/month using the simplistic Amazon Glacier calculator[80]. Yet, such tools do not model the full cost of open data sharing: cloud storage options often carry costs for data retrieval on top of storage, ranging from pennies to $0.090 per GB/month. At the other end of the spectrum is the Curation Costs Exchange (CCEx)[81], a community-owned platform for cost modeling that includes options for modeling finer-grained discovery and labor-related costs in addition to storage, and also supports output comparison with crowd-sourced data from peer institutions. This is a useful tool for modeling the basic costs of digital archiving based on length of preservation term, bytes stored, and other factors including labor. With some seed funding cost models for MPS could be incorporated in or extended from CCEx to support specific cost scenarios for MPS data which may have more complex digital objects, large file sizes, and dependencies on software components which must be preserved alongside data to support re-use.

Getting to 2030: Open Access as the norm

It is useful to discuss the notions of Open Access described in this report in a scenario of relatively unlimited resources to understand where this might lead. As mentioned in the previous sections, a globally discoverable and searchable data archive could be layered with the knowledge required for re-use, increasing the usefulness of the stored data over time as more information and more tools are added. One could imagine queries of the scholarly record such as the following becoming routine:

- List all the image denoising algorithms used on Pandora's galaxy cluster Abell 2744, and the input parameters used, in the last 5 years;
- Find all the input data used in predictions of Hurricane Katrina's approach path, prior to its landing;
- Give the complete set of code, parameters, data, and workflows used in the recent successful detection of Gravitational Waves;
- Create a unified dataset of adsorption capacities of materials within a 3 mile radius of the Palo Verde 1 Nuclear Power Plant;
- Create a unified dataset of all published whole-genome sequences identified with mutation in the gene BRCA1[82]
- Identify a newly synthesized compound by matching collected new data points

---

[80] Amazon Glacier: https://aws.amazon.com/glacier/pricing/

[81] CCEx: http://www.curationexchange.org/

[82] Gavish, M & Donoho D, "Three Dream Applications of Verifiable Computational Results", Computing in Science & Engineering vol. 14 no. , p. 26-31 , 2012; https://www.computer.org/csdl/mags/cs/2012/04/mcs2012040026-abs.html

against a repository of all data for known compounds.
- Locate a material that matches specific property needs (e.g., high conductivity, malleable) searching across a centralized physical property database

Such queries of the scholarly record are not easy today, and in many cases impossible. Yet we argue that not having the infrastructure in place to routinely make such queries – that depend crucially on persistent links between published results, underlying data and software, and open access – limits how the scientific community can make important discoveries and understand the existing literature.

We also believe verification of experimental and computational results will become routine for the majority of published articles. One can imagine an automated check that executes codes to verify that these data with this analysis produced these findings. Of course, this would not ascertain the scientific correctness or value of the results, but it would ensure the computational record was transparent, testable, and able to be inspected, as long as the capability for computation is appropriately preserved. Additionally, this would enable a discovery environment that would permit the comparison of a method on different datasets, and the comparison of different methods on a unified dataset. Furthermore, aggregation and federation of distributed data enables new and previously unexplored discovery spaces as well as improving existing capabilities for cross comparison and meta-analysis.

Pilot projects

The workshops discussed several concrete proposals for pilot projects or programmatic initiatives that could serve as small steps toward the construction of an open-access data ecosystem. Some of these are merely simple ideas, others are larger projects that could produce building blocks of a larger system.

- Certified repositories:
    - Support creation of "advanced" repository systems that can ingest the broad spectrum of data associated with knowledge preservation
    - Curate lists of certified archives and their uses
    - Inreach to the scientific communities in order to
        - Publicize the capabilities and uses of new repositories, such as embargo capabilities, cross-platform data sharing and computation, etc.
        - Initiate discussion of standards

- Establish prototype federated archival systems:
    - Create interoperable links between disparate resources, such as
        - National Data Service
        - Regional data repositories
        - Large subject independent repositories such as figshare, Dryad
        - University repositories,

        - Domain-specific repositories (e.g., CERN, NASA, NIST)

- Attach additional funding to grants, or have separate RFPs that encourage more effective modes of work in terms of data/knowledge preservation

- Pilot projects to demonstrate benefits of workflow preservation, use of data management tools, etc.
    - e.g., development of sophisticated electronic log books that can capture workflows and data; using this to share results
    - Develop shared projects between communities or societies representing those communities and funding agencies such as NSF. Seek to develop model workflows for submission, review/verification, publication, and preservation of specific data types with a focus on resulting in new use cases that advance science

- Tools for automatic metadata generation
    - Combine metadata, computer science experts, etc. to arrive at generic capabilities for metadata generation based on workflow/processing tools

- Metadata development:
    - Develop searchable and computable ontologies for knowledge preservation, including workflows, multiple data sources, etc.

- Development of training materials for data and workflow preservation tools
    - workforce development will be an important aspect of these efforts, since they represent a new way of doing science. Achieving acceptance and implementation at a grass roots level will be crucial for changing the research culture.

- Study of use cases and gap analysis for supporting data flow through the full research cycle specifically for the MPS domains; to identify additional challenges beyond the technical, and potential partnerships with other stakeholders to facilitate greater adoption of depositing data

In addition to a consideration of how some of the pilot projects suggested above might be incorporated into the strategic plans of the agency, we arrived at several proposals as to how the NSF might accelerate the creation of the open access data ecosystem. Two of these seem sufficiently important to mention them here:

- The MPS community needs examples of excellence in order to lower the barriers to preparing scientific results for sharing and archiving. NSF could highlight and disseminate best practices for data preservation and sharing. This could take the form of encouraging publication of excellent examples of Data Management Plans, showcasing the state of the art in published datasets with

discoverable products, highlighting important scientific results derived from re-use of public datasets, etc. Having the DMPs linked to abstracts and the data or other products resulting from the grant would be an excellent resource for researchers seeking to follow these exemplars.

- The MPS community needs guidelines for trusted repositories in line with the NSF policy on disseminating research[83]. NSF does not have the infrastructure to store all of the research data that scientists may want to archive. Researchers have available to them a variety of repositories of varying quality and sophistication, not all of which meet ISO 16363 standards for trusted repositories. Some minimum set of requirements should be established so that MPS scientists know that they have stored their data in a repository that meets the standards for a trusted repository as far as NSF policy is concerned. These guidelines should address, at a minimum, such concerns as data security, licensing, and the quality of bit-level integrity checking.

Conclusions

This report attempts to summarize current attitudes and practices of the MPS science communities regarding the preservation, sharing, and re-use of the knowledge associated with the production of scientific results. It is a result of two workshops that brought together thoughtful representatives of the various MPS communities and extensive outreach to gather responses to the proposed conclusions. There is broad but certainly not unanimous support for public access to research data. Herein, we have outlined initial steps toward public access to research results and supporting documentation that are generally acceptable to the MPS research communities. From the perspective of the needs of MPS research scientists, we have also sketched a roadmap for achieving open access for re-use and reproducibility of scientific results. The effort and investment to achieve that goal, if desired, will be substantial. Not only will a much more extensive cyberinfrastructure be required to support the technical aspects of preservation and re-use, but broad changes to the way science is done and the culture that surrounds it will need to take place. While many of us believe that motion toward open access would be beneficial for science, it is clear that this goal will be difficult to attain. We hope that the considerations presented here can serve as a guide to those contemplating the construction and coordination of an open access future.

Respectfully submitted:

Workshop PIs and Organizers:

Ani Aprahamian,
Tim Beers,

[83] https://www.nsf.gov/pubs/manuals/gpm05_131/gpm7.jsp - 734

Steve Buechler,
Mike Hildreth,
Jarek Nabrzyski

Report Editors:

Robert Hanisch (NIST),
Michael Hildreth (Notre Dame),
Leah McEwen (Cornell),
Victoria Stodden (UIUC),
Gordon Watts (UW),
Daniel S. Katz (UIUC),
Natalie Meyers (Notre Dame),
Ashley E. Sands (UCLA)

Appendix I: Funding and Cost Model Calculators and Resources

The following resources were considered as possible reference models for calculating the cost of knowledge preservation infrastructure. A brief overview of each tool is presented below.

- AWS Cost Calculator: http://calculator.s3.amazonaws.com/index.html
- Amazon Glacier Pricing: https://aws.amazon.com/glacier/pricing/
- Azure Calculator: https://azure.microsoft.com/en-us/pricing/calculator/
  ~20GB @ 1.60 per month = 19.20 yearly

Curation Costs Exchange (CCEx) (http://www.curationexchange.org/)
The Curation Costs Exchange (CCEx) is a community owned platform which helps organizations of any kind assess the costs of curation practices through comparisons and analysis. It produces graphs like the one shown in Figure I.1.:

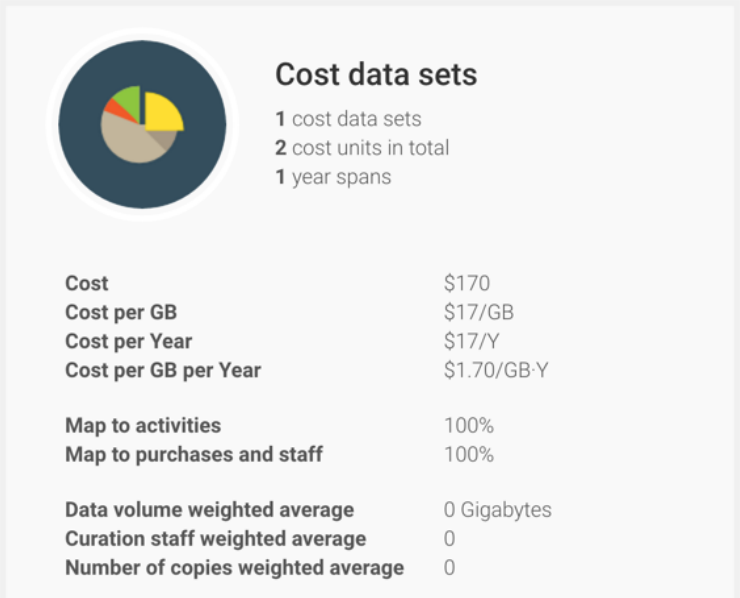


Figure I.1 Example output from CCEx for a given dataset.

Figure I.2. A comparison of data curation costs across peer institutions, from the CCEx platform.

CCEx also allows comparison with peer institutions' data, as shown in Figure I.2.

The 4C project carried out a review[84] of ten existing curation and preservation cost models. The models can be accessed at this url: http://www.4cproject.eu/summary-of-cost-models. A summary of the models considered is shown in Table I.1.

ID	Name	Acronym	Owner
1	Test bed Cost Model for Digital Preservation	T-CMDP	National Archives of the Netherlands
2	NASA Cost Estimation Tool	NASA-CET	National Aeronautics & Space Administration (NASA)
3	LIFE^3 Costing Model	LIFE3	University College London and the British Library
4	Keeping Research Data Safe	KRDS	Charles Beagrie Ltd
5	Cost Model for Digital Archiving	CMDA	Data Archive and Networking Services (DANS)
6	Cost Model for Digital Preservation	CMDP	Danish National Archives and the Danish Royal Library
7	DP4LIB Cost Model	DP4LIB	The German National Library
8	PrestoPRIME Cost model for Digital Storage	PP-CMDS	The PrestoPRIME Project
9	Total Cost of Preservation	CDL-TCP	California Digital Library
10	Economic Model of Long-Term Storage	EMLTS	David Rosenthal

Table I.1. The cost models studied by the 4C consortium. See references in text.

The NASA Cost Estimation Tool (NASA CET)[85].
The NASA Cost Estimation Tool produces reports like this which model the fully loaded cost of labor associated with data preservation from ingest, to back up, to retrieval, and support as well as the cost of physical storage. The following information can be found on the NASA CET web site.

[84] http://www.dcc.ac.uk/projects/4c/cost-model-comparison-table
[85] https://opensource.gsfc.nasa.gov/projects/CET/

NASA CET Purpose: Estimating life cycle costs for ground data centres activated to improve budgets for NASA missions.
Creator and Funding: Developed for NASA by SGT (Stinger Ghaffarian Technologies Inc.).
Status: The first version of CET (version 1) was published in 2004. The newest version available (version 2.4) is from September 2008.
Activities Modeled: Ingest, Data Management, Archival Storage, Access, Administration.
Resources: 96 distinct descriptors such as staff salaries, system purchase cost, COTS software license, archive media, inflation, volume, automation level.
Time Period: Past, Present, Future - for a 7 to 10 year time scale reflecting normal data processing time period for missions.
Variables: Labour salaries (6 types),capital cost (building space, hardware and software, clients, servers, databases, storage, 'archive system'), migration frequency, number of assets.
Type of tool: Analysis, estimation and review of estimation. Implemented in MS Excel Spreadsheet using Visual Basic (VBA).
Availability of tool: The CET tool, user guide, technical description etc is available for download at http://opensource.gsfc.nasa.gov/projects/CET/ (zip archive at bottom of page: http://opensource.gsfc.nasa.gov/projects/CET/CET%20V2.4.zip).

Table I.2 shows some example output for a trial dataset, giving an impression of the level of detail produced by the tool.

Revised Estimate	Activity Dataset:	OpenMPSTest1							Produced:	12/01/16				
Mission Start Year:	2017	Operations Start Year:	2017		Mission Complete Year:	2026		Inflation Rate	3.0%					
Estimated Staffing Level	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	Total	Pct.
Management Staff FTE	1.55	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.00	0.00	10.32	8.3%
Administrative Support FTE	0.30	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.00	0.00	2.03	1.6%
Technical Coordination Staff FTE	1.25	1.25	1.25	1.25	1.25	1.25	1.25	1.25	1.25	1.25	0.00	0.00	12.50	10.1%
Development Staff FTE	5.56	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.56	4.5%
Technical / Science Staff FTE	4.69	4.69	4.69	4.69	4.69	4.69	4.69	4.69	4.69	4.69	0.00	0.00	46.90	37.8%
Operations Staff FTE	3.02	3.02	3.02	3.02	3.02	3.02	3.02	3.02	3.02	3.02	0.00	0.00	30.22	24.3%
Sustaining Engineering Staff FTE	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.00	0.00	8.44	6.8%
Engineering Support Staff FTE	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.00	0.00	8.25	6.6%
Estimated Total FTE	18.05	11.80	11.80	11.80	11.80	11.80	11.80	11.80	11.80	11.80	0.00	0.00	124.21	
Estimated Staff Costs, K$	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	Total	Pct.
Management Staff Cost	16	10	10	11	11	11	12	12	12	13	0	0	117	4.9%
Administrative Support Staff Cost	3	2	2	2	2	2	2	2	2	2	0	0	23	1.0%
Technical Coordination Staff Cost	13	13	13	14	14	14	15	15	16	16	0	0	143	6.0%
Development Staff Cost	278	0	0	0	0	0	0	0	0	0	0	0	278	11.6%
Technical / Science Staff Cost	47	48	50	51	53	54	56	58	59	61	0	0	538	22.4%
Operations Staff Cost	30	31	32	33	34	35	36	37	38	39	0	0	346	14.5%
Sustaining Engineering Staff Cost	42	43	45	46	48	49	50	52	53	55	0	0	484	20.2%
Engineering Support Staff Cost	41	42	44	45	46	48	49	51	52	54	0	0	473	19.7%
Total Estimated Staff Cost	470	189	196	202	208	213	220	227	232	240	0	0	2,397	49.6%
System Purchase Cost	2	0	0	0	0	0	0	0	0	0	0	0	2	0.1%
COTS Software License Cost	22	2	2	2	2	2	2	2	2	2	0	0	40	1.6%
Facility Preparation and Support Cost	374	177	177	177	177	177	177	177	177	177	0	0	1,967	80.8%
System Maintenance Cost	4	4	4	4	4	4	4	4	4	4	0	0	40	1.6%
Network / Communications Cost	4	4	4	4	4	4	4	4	4	4	0	0	40	1.6%
General Supplies Cost	37	28	28	28	28	28	28	28	28	28	0	0	289	11.9%
Archive Media Cost	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0%
Distribution Media Cost	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0%
Travel Cost	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0%
Training Cost	20	4	4	4	4	4	4	4	4	4	0	0	56	2.3%
Data Purchase Cost	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0%
Computer Services Cost	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0%
Total Estimated Non-Staff Costs, K$	463	219	219	219	219	219	219	219	219	219	0	0	2,434	50.4%
Total Estimated Cost, K$	933	408	415	421	427	432	439	446	451	459	0	0	4,831	

Table I.2. Sample output of the NASA CET using inputs from a trial dataset.

Related Reading:

"Funding research data management and related infrastructures" Authors: Knowledge Exchange Research Data Expert Group and Science Europe Working Group on Research Data May 2016 .

Han, Y. (2015). Cloud storage for digital preservation: Optimal uses of amazon S3 and glacier. *Library Hi Tech, 33*(2), 261-271. Retrieved from http://proxy.library.nd.edu/login?url=http://search.proquest.com.proxy.library.nd.edu/docview/1684437549?accountid=12874 or here in drive during workshop.

> Findings - Cloud storage solutions like Glacier can be very attractive for long-term digital preservation if data can be operated within the provider's same data zone and data transfer-out can be minimized. Practical implications - Institutions can benefit from cloud storage by understanding the cost models and data retrieval models. Multiple strategies are suggested to minimize the costs. Cloud storage pricing especially data transfer-out pricing charts are presented to show the price drops over the past eight years.
>
> "the big three's pricing is currently set at USD 0.03/ GB, a 60 percent of drop from 0.085/ GB in January 2014. S3's data transfer-in is now totally free, and data transfer-out has also dropped from 0.12/ GB (2011) to 0.09/ GB (2014). "

Report on cost parameters for digital repositories Project: APARSEN Doc. Identifier: APARSEN-REP-D32_1-01-1_0 Date: 2013-02-28 D32.1
High level analysis of published cost models as well as the initial findings of a review of their cost parameters is provided.

Report on testing of cost models and further analysis of cost parameters Date: 2013-06-30 Project: APARSEN Doc. Identifier: APARSEN-REP-D32_2-01-1_0
The results of the analysis of cost parameters and the testing of cost models used within digital preservation both for services and repositories is provided. The relationship between costs and benefits is also reviewed in the context of digital preservation.

David S.H. Rosenthal, Daniel C. Rosenthal, Ethan L. Miller, Ian F. Adams, Mark W. Storer and Erez Zadok. "The Economics of Long-Term Digital Storage", Memory of the World in the Digital Age, Vancouver, BC, September 2012.

Report on Knowledge Exchange Workshop: Cost models for keeping knowledge: economic models for digital preservation, 11 June 2012

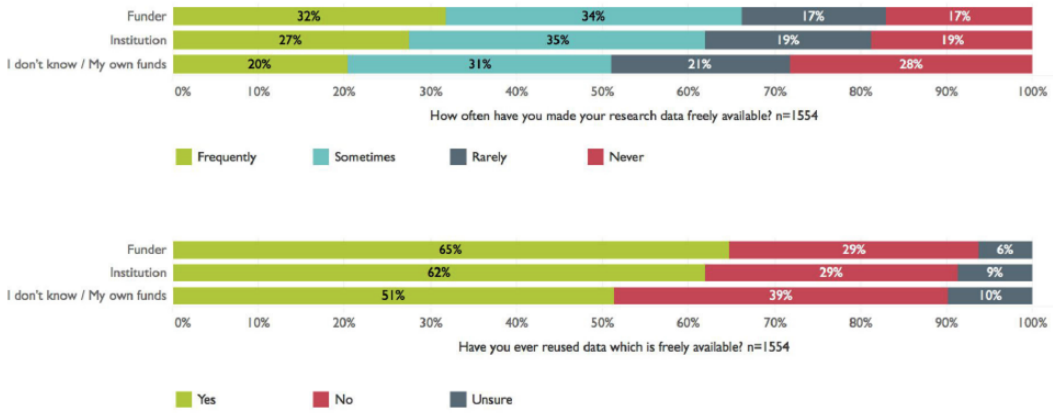
Preserving Scientific Data on Our Physical Universe:
A New Strategy for Archiving the Nation's Scientific Information Resources(1995)
https://www.nap.edu/catalog/4871/preserving-scientific-data-on-our-physical-universe-a-new-strategy

Vendor: http://alinean.com/ creates custom cost calculators for industry.

Who Pays?

A question related to funding from the figshare "State of Open Data" Report: who do researchers think should pay for the storage of persistent data?

Figure H - Who would meet the costs of making your research data openly available?



Treadway, Jon; Hahnel, Mark; Leonelli, Sabina; Penny, Dan; Groenewegen, David; Miyairi, Nobuko; Hayashi, Kazuhiro; O'Donnell, Daniel; Science, Digital; Hook, Daniel (2016): ***The State of Open Data Report***. Figshare.
https://dx.doi.org/10.6084/m9.figshare.4036398.v1

This Appendix content originally prepared by Natalie K. Meyers, E-Research Librarian, University of Notre Dame, for participants of the Open MPS Workshop on "Gauging the Impact of Requirements for Public Access to Data" Arlington, VA Dec 1-2 2016
https://mpsopendata.crc.nd.edu/index.php/w-2/about-w2
The MPS Open Data workshop series is supported by the National Science Foundation.

Appendix II: Contents of APS Survey

The following questions were included in the survey submitted by the APS to over 5000 researchers who are authors in APS journals.

1. In the past three years, have you or your research group made publicly accessible the following items through your or your institution's website or a third-party repository?

Answer Options	Yes	No	Unsure	N/A
Raw data				
Structured databases				
Processed data				
Figure/Plot/Table data				
Software				

2. Do you or your research group have an established practice for archiving the Do you or your research group have an established practice for archiving the following items?

Answer Options	Yes	No	Unsure	N/A
Raw data				
Structured databases				
Processed data				
Software				
Software environment				
Analysis workflow				

3. In your estimation, which of the following currently have the infrastructure required to provide long-term public access to your research data?

Answer Options	Yes	No	Unsure
Your research group			
Your institution			
Your funding agency			
Third-party repositories			
Journal publishers			

4. Ignoring software and documentation, what is the typical storage space required by your research group to store the following kinds of data?

Answer Options	<1 MB	1 - 10 MB	10-100 MB	100 MB - 1 GB
Raw data				
Structured databases				
Processed data				
Figure/Plot/Table data				

Answer Options	1 - 10 GB	10 - 100 GB	100 GB - 1 TB	1 - 10 TB
Raw data				
Structured databases				
Processed data				
Figure/Plot/Table data				

Answer Options	10 - 100 TB	> 100 TB	Unsure	N/A
Raw data				
Structured databases				
Processed data				
Figure/Plot/Table data				

5. Please indicate the level of staffing/funding that would be required for you or your research group to make publicly accessible the following items on a sustained basis.

Answer Options	Already doing it	Could be done with existing staff/funding	Would need additional staff/funding	Not practicable	N/A
Raw data					
Structured databases					
Processed data					
Figure/Plot/Table data					
Software					

6. Please indicate the level of your agreement or disagreement with the following statements.

Answer Options	Strongly disagree	Somewhat disagree	Unsure	Somewhat agree	Strongly agree
My research has benefited from the current practices of data and software sharing in my field.					
I have sufficient access to underlying data, software, and documentation in order to reproduce, reuse, and/or validate experimental or theoretical results in my field.					

7. If the necessary infrastructure and funding were available, how inclined would you be to make your data, software, and documentation necessary for its interpretation publicly available?

Answer Options
I already make my data publicly available.
I would endeavor to make more of my data publicly available
I'd like to make my data publicly available, but it would be too time consuming or expensive to prepare it for use by others
I would not be inclined to make my data publicly available.

8. In the past three years, have you or your research group privately shared data and/or software with other researchers or research groups?

Answer Options
Yes
No
Unsure

9. In the past three years, has another researcher or research group privately shared their data and/or software with you or your research group?

Answer Options
Yes
No
Unsure

10. In the past three years, have you published a journal article that cites data made publicly available by another research group?

Answer Options
Yes
No
Unsure

11. In the past three years, has another research group published a journal article citing data made publicly available by you or your research group?

Answer Options
Yes
No
Unsure

Demographic questions, and an open-ended opportunity for comment rounded out the survey.

Appendix III: Additional Surveys of Interest to the MPS Community

The provides a snapshot of open data sharing practices and a timely complement to a previous study: ***Science Gateways Today and Tomorrow*: Positive perspectives of nearly 5,000 members of the research community**.

Taken together these two surveys paint a broad picture of those producing, reusing, and making research data more open. 74% of *State of The Data* survey respondents have made research data open at some point, and of respondents who have never done so, 90% would consider making data open in the future. This interest closely echoes the Science Gateways survey, where 75% of respondents indicated that data collections were important to their research/education work, ranking it highly alongside data analysis tools and computational tools (72% each) and their interest in being able to rapidly publish and/or find domain-specific articles and data (69%).

Note that Science Gateways' survey respondents have a disciplinary overlap with the Open MPS audience, as shown in the figure, below.

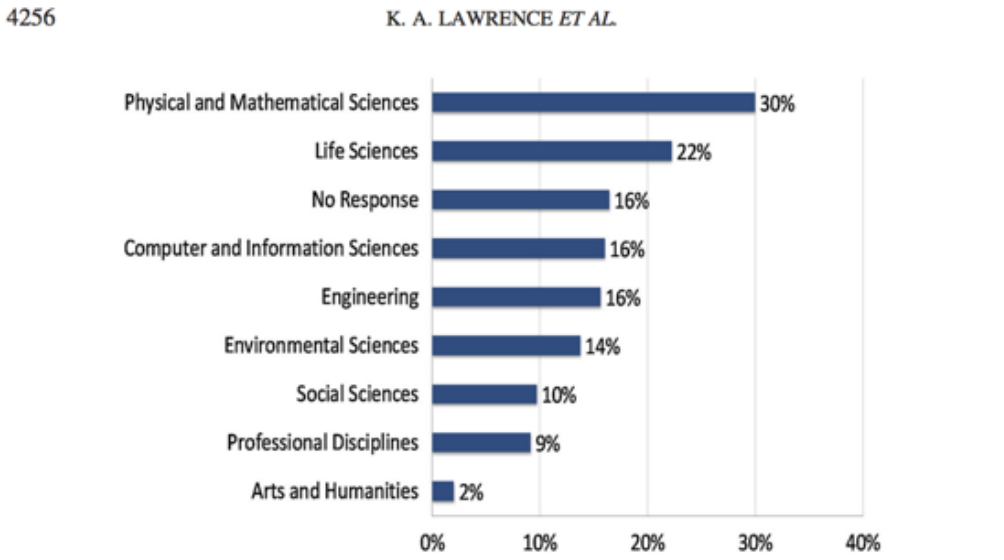


Figure 1. Primary areas of current domain expertise. Respondents could select all that apply; 84% (4141 of 4957 participants) responded, generating 6689 total responses (mean = 1.6 domains per respondent). 'Professional Disciplines' included Architecture; Education; Information; Law, Legal Professions, and Studies; Library Science; Management or Business; Natural Resources, Forestry, and Conservation; Public Administration; Social Service Professions; and 'other'.

References:

Treadway, Jon; Hahnel, Mark; Leonelli, Sabina; Penny, Dan; Groenewegen, David; Miyairi, Nobuko; Hayashi, Kazuhiro; O'Donnell, Daniel; Science, Digital; Hook, Daniel (2016): ***The State of Open Data Report***. Figshare. https://dx.doi.org/10.6084/m9.figshare.4036398.v1

(NPG), Nature Publishing Group (2016): ***Open Data Survey***. figshare. https://doi.org/10.6084/m9.figshare.4010541.v4

Lawrence, K. A., Zentner, M., Wilkins-Diehr, N., Wernert, J. A., Pierce, M., Marru, S., and Michael, S. (2015) ***Science gateways today and tomorrow: positive perspectives of nearly 5000 members of the research community***. Concurrency Computat.: Pract. Exper., 27: 4252–4268. doi: 10.1002/cpe.3526.

A recently-released survey of attitudes and practices related to data preservation and sharing was released in April 2017. A partnership between Elsevier Publishing and the University of Lieden, the survey covered many of the same topics as those reported above, as well as the APS survey conducted as part of this workshop series.
Many of the same rates of sharing and attitudes toward open data as have been found previously were also reported in this survey.

Reference:
Berghmans, Stephane; Cousijn, Helena; Deakin, Gemma; Meijer, Ingeborg; Mulligan, Adrian; Plume, Andrew; de Rijcke, Sarah; Rushforth, Alex; Tatum, Clifford; van Leeuwen, Thed; Waltman, Ludo (2017), "Open Data: the researcher perspective - survey and case studies". Mendeley Data, v1
http://dx.doi.org/10.17632/bwrnfb4bvh.1

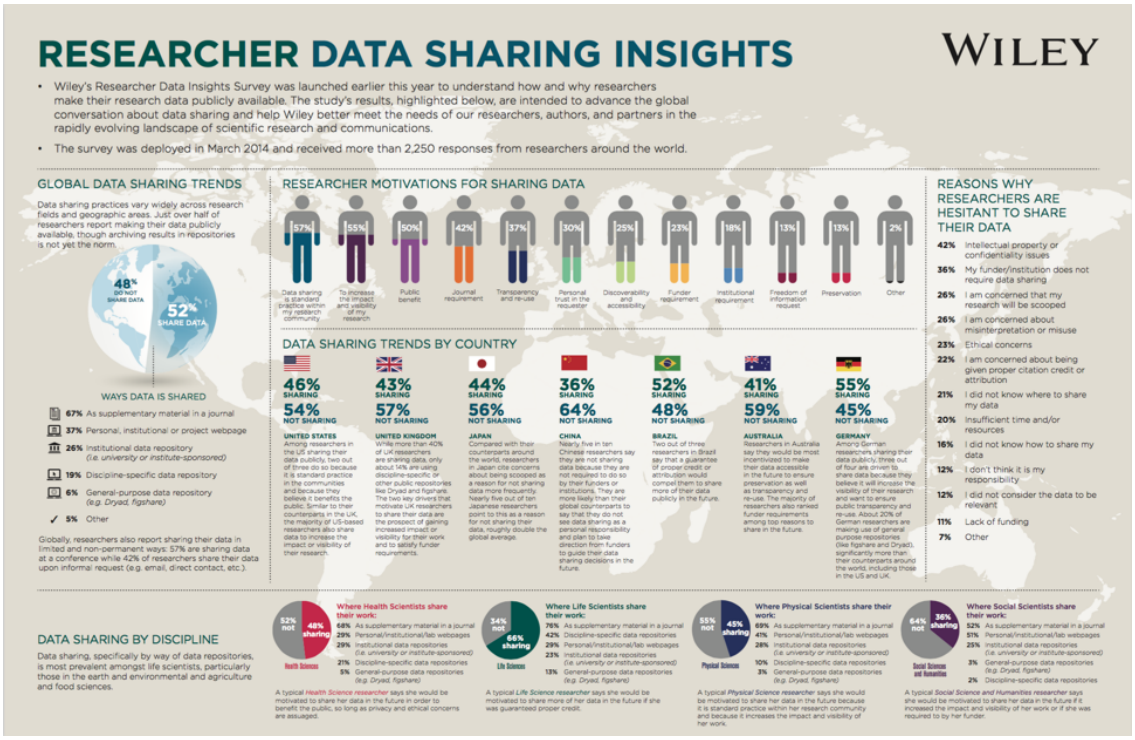
The ***Data and Software Preservation Quality Tool Planning Project (PRESQT)*** will be conducting a complementary survey in 2017. Open MPS workshops participants and other interested parties are encouraged to contact PRESQT (http://presqt.crc.nd.edu/) if there are questions it is important to repeat or new questions that should be included in the forthcoming survey to better inform preservation and access to Open Math and Physical Sciences data.

Austrian National Data Survey:
From 19th to 31st January 2015, all scientists associated with the 25 Partners of e-Infrastructures Austria were invited to participate in the National Research Data Survey.

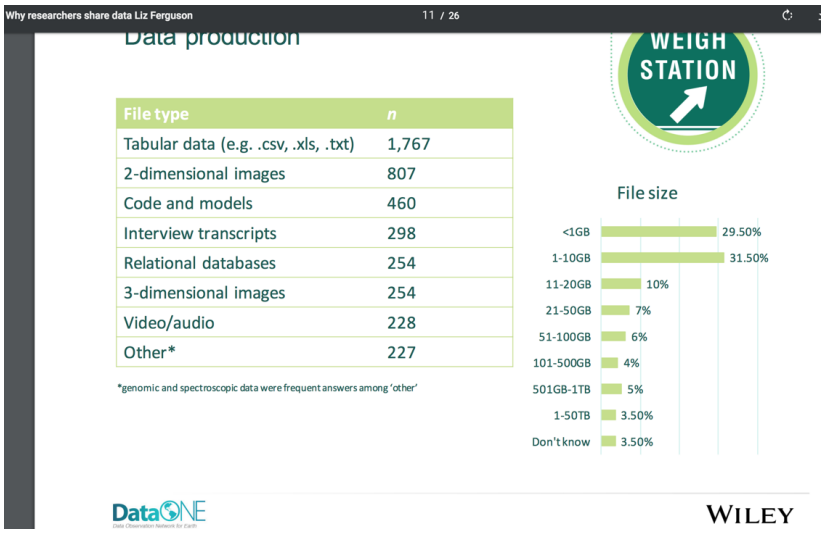
- Blog post and infographic: https://blogs.openaire.eu/?p=619
- Full report: https://zenodo.org/record/34005
- Full questionnaire is here: https://e-infrastructures.at/fileadmin/user_upload/p_e_infrastructures/PDFs/Questionnaire_e-Infra_Research_Data_Survey_Aug15.pdf

Wiley Researcher Data Insights Survey was conducted 2014, data posted to Figshare in Aug 2016(https://doi.org/10.6084/m9.figshare.3468368.v2).

This Wiley Data Sharing Project was undertaken to understand how and why researchers make their research data publicly available. This infographic (below) of results has been shared:



And, some of the Wiley survey responses were described in Liz Ferguson's presentation for DataOne "How and Why Researchers Share Data" Nov 2015. Of interest to MPS effort may be the self-reported file sizes and data production formats.



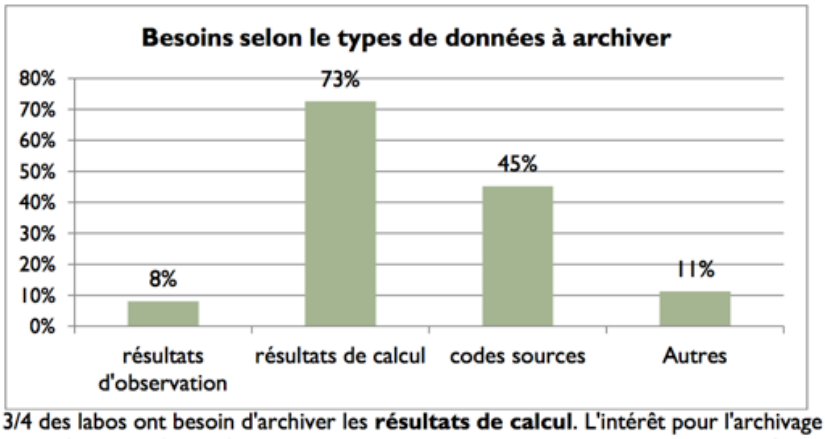
Van den Eynden, V. and Bishop, L. (2014). **Incentives and motivations for sharing research data, a researcher's perspective**. A Knowledge Exchange Report, available from knowledge-exchange.info/Default. aspx?ID=733

Materials Research Society (MRS) and The Minerals, Metals and Materials Society (TMS) 25-question "***MRS-TMS Big Data Survey***." 2013 See: Wilson, L. (2013). Survey on Big Data gathers input from materials community. *MRS Bulletin, 38*(9), 751-753. doi:10.1557/mrs.2013.225

There were 675 respondents when the survey was closed on June 3, of which 73% completed the survey. The other 27% responded to most of the questions but did not finish the survey.

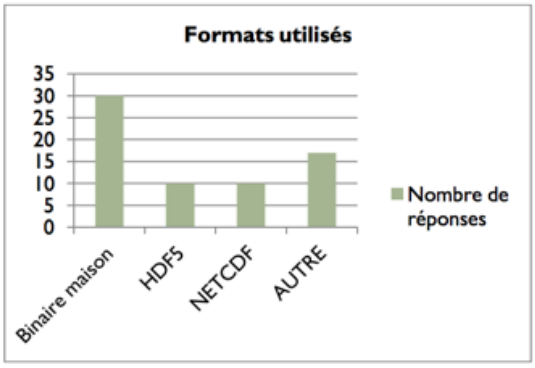
> The top three motivations that survey respondents cited as encouragements for sharing their data on an open-access basis were (1) increased visibility of research/work (72%), (2) the opportunity to receive feedback from others about the data (67%), and (3) the opportunity for others to analyze the data (and potentially make other discoveries as a result) (54%). Conversely, the top impediments identified by survey respondents were (1) the proprietary/ restricted nature of their data (59%), (2) the intellectual property rules within their organization/business (54%), and (3) the fact that their data was stored in a proprietary data format (42%).
>
> "Ultimately, the goal is better processes and tools for sharing information to advance the overall field of materials."

ISAAC (Scientific Information Archived At CINES) Survey: This survey was conducted in June 2011. It was not only a question of CINES' users with regard to the archiving of their data, but also to evaluate the knowledge and practices of the community in the field. Nombre de questionnaires envoyés: 155 Nombre de réponses (complètes ou incomplètes): 62 (40%), dont 38 complètes (25%).



3/4 des labos ont besoin d'archiver les **résultats de calcul**. L'intérêt pour l'archivage des codes sources est partagé, la moitié des réponses indiquent que les **labos peuvent gérer eux même** cet archivage car les volumes sont limités. Peu de demandes pour l'archivage des résultats d'observation, à relativiser avec le nombre de projets qui font ce type de mesures au CINES.
Les autres types de données : les **données en entrée des codes**, les **résultats d'analyse**, les **visualisations**, les **sources des publications** (simulations publiées).

➢ **Les formats utilisés:**



Une majorité des projets ont des données au **format binaire**, en sortie des logiciels ou des codes de calcul utilisées. **Les fichiers ASCII** et **textes** sont utilisés dans 1/3 des projets souvent en compléments des données calcul. Les données en **HDF5** et **NETCDF** sont aussi très présentes.
D'autres formats sont utilisés plus rarement mais peuvent être intéressants à étudier (FITS, Grib, CGNS).

NIPS Survey. Victoria Stodden conducted a Web-Sharing Practices in Computational Research survey. She questioned researchers registered for Neural Information Processing Systems (NIPS) conference, held annually in Whistler, British Columbia, Canada

Stodden, Victoria, The Scientific Method in Practice: Reproducibility in the Computational Sciences (February 9, 2010). MIT Sloan Research Paper No. 4773-10. Available at SSRN: https://ssrn.com/abstract=1550193 or http://dx.doi.org/10.2139/ssrn.1550193

See also the web questionnaire for above NIPS survey :
http://web.stanford.edu/~vcs/Survey2009/SharingSurvey.html

This Appendix originally sourced and prepared by Natalie K. Meyers, E-Research Librarian, University of Notre Dame for participants of Open MPS Workshop on "Gauging the Impact of Requirements for Public Access to Data" Arlington, VA Dec 1-2 2016 https://mpsopendata.crc.nd.edu/index.php/w-2/about-w2

The MPS Open Data workshop series is supported by the National Science Foundation.

Appendix IV: Workshop 1 Registrants

Allen	Gale	NASA
Aprahamian	Ani	University of Notre Dame
Beers	Timothy	University of Notre Dame
Boehm	Reid	University of Notre Dame
Buechler	Steve	University of Notre Dame
Chalk	Stuart	University of North Florida
Davies	Kevin	American Chemical Society
de Waard	Anita	Elsevier
Falk-Krzesinski	Holly	Elsevier
Garritano	Jeremy	University of Maryland
Hanisch	Robert	NIST
Hildreth	Mike	University of Notre Dame
Hunter	Angie	Organic Letters (ACS Publication)
Juric	Mario	University of Washington
Katzgraber	Helmut G	Texas A&M University
Kliemann	Wolfgang	Iowa State University
Langston	Glen	National Science Foundation
Lauret	Jerome	Brookhaven National Laboratory
Lewis	Alexis	National Science Foundation
Martinsen	David	American Chemical Society
McEwen	Leah	Cornell University
Meyers	Natalie	University of Notre Dame
Nabrzyski	Jarek	University of Notre Dame
Olsen	Karen	NIST
Petravick	Donald	NCSA/University of Illinois
Phillips	Mary	Oregon State University
Proffen	Thomas	Oak Ridge National Laboratory
Robinson	Carly	Department of Energy
Savage	Martin	University of Washington
Sharp	Nigel	National Science Foundation
Smale	Alan	NASA Goddard Space Flight Center
Stahlman	Gretchen	University of Arizona
Steffen	Julie	American Astronomical Society
Stodden	Victoria	University of Illinois Urbana-Champaign
Thakar	Ani	The Johns Hopkins University
Watts	Gordon	University of Washington

Appendix V: Workshop 2 Registrants

Allen	Gale	NASA
Biven	Laura	Department of Energy
Bloemhard	Heather	American Astronomical Society
Boehm	Reid	Johns Hopkins University
Boswell	Joshua	C&M International
Chalk	Stuart	University of North Florida
Chiu	Chen	Johns Hopkins University
de Waard	Anita	Elsevier Publishing
Donaldson	Devan	Indiana University
Gaal	Rachel	APS
Garritano	Jeremy	University of Virginia
Goroff	Daniel	A.P. Sloan Foundation
Hanisch	Robert	NIST
Hanson	Brooks	American Geophysical Union
Henderson	Darla	ACS
Henneken	Edwin	Harvard University
Hu	Allen	APS
Huck	Patrick	Lawrence Berkeley Laboratory
Katz	Daniel	University of Illinois Urbana-Champaign
Knezek	Patricia	NSF
Lewis	Alexis	NSF
Martinsen	David	Martinsen Consulting
McEwen	Leah	Cornell University
Pasquetto	Irene	UCLA
Potterbusch	Megan	Association of Research Libraries
Robinson	Carly	Department of Energy
Sands	Ashley E.	UCLA
Stahlman	Gretchen	University of Arizona
Stall	Shelley	American Geophysical Union
Steffen	Julie	American Astronomical Society
Stodden	Victoria	University of Illinois Urbana-Champaign
Tein	Andrew	Wiley Publishing
Ward	Charles	USAF Research Laboratory
Warren	James	NIST
Watts	Gordon	University of Washington
Weinreich	David	Weinreich Strategic Group