MPS Open Data Workshop 1:

Draft Report[1]

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 citizen scientists and other experts, perhaps enhancing the potential for further discovery. As an additional benefit, an effective implementation of public access to data, interpretive tools, and results could be a critical ingredient in restoring public trust in the scientific enterprise since these are essential components for insuring the reproducibility of results.

As of 2013, it is also now federal policy[2] that results and data derived from publicly-funded research be made available to the public at large. After the announcement of the new policy by the Office of Science and Technology Policy at the Whitehouse, each federal agency was required to develop an implementation plan stating how this directive would be followed. The National Science Foundation implementation plan, "Today's Data, Tomorrow's Discoveries" (NSF 15-52),[3] was released in March 2015. The report outlines a staged strategy for increasing public access to research results, beginning with enabling access to published journal articles and juried conference submissions. It leaves in place the current policies of the Data Management Plans (DMPs) related to the conservation and sharing of research data, but includes a clear statement of motion towards 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."[4]

This report is a direct result of consultation with the research communities funded by the Directorate of the Mathematical and Physical Science (MPS) at NSF. Sponsored by an NSF award, the first of two planned workshops was held in Arlington, VA, on November 19 and 20, 2015, to "take the pulse of the research community" on public access to research data. The goal of this effort is to provide feedback to NSF on current best practices with regard to research data curation and access, and suggestions for areas of improvement and investment that could facilitate broader access to research data in the future. The ideas that emerged from the first workshop have been captured in the text below so that they can be shared

---

[1] Primary authors: Robert Hanish (NIST), Michael Hildreth (Notre Dame), Leah McEwen (Cornell), Victoria Stodden (UIUC), Gordon Watts (UW)

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

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

[4] ibid, p. 9.

with the broader community of MPS-funded researchers. The eventual outcome will be a report representing input from a wide community of MPS scientists. A second workshop will be convened in December 2016 to accept the input collected over the next few months and to produce the final report.

Executive Summary

This draft report contains a broad discussion of open access to data and what this might imply for the NSF-funded researcher within the MPS directorate. Broadly speaking, the conclusion of the deliberations is that the research culture, the data management tools, and the archive infrastructure are not ready at this time for a move to full access to all research data. The report does make one concrete conclusion for immediate implementation:

Data upon which publications are based should be available in machine-readable digital format, and persistently linked to those publications.

This relatively simple suggestion is one in a series of steps outlined as a roadmap towards achieving a goal of full re-use of research data, and one that would already make a large difference in transparency and credibility for research results coming from MPS. The bulk of the report is a discussion of the appropriate elements that would characterize a functional open access data infrastructure. These include a description of the aspects of a research project that should be stored along with and linked to the data and publication if full re-use is the desired outcome:

- **Software:** the software used to create, process, and analyze the data
- **Workflow:** instructions, frameworks, or scripts use to run the software
- **Software environment:** a specification or an instantiation of the requisite operating system, architecture, libraries, machine state, etc., that are necessary to run the software/workflows
- **Simulation capabilities:** the capability to run the software with different parameters than used to generate the original data
- **Documentation:** a description of the software, workflows, and other information describing how the data were derived, processed, and analyzed.
- **Data characterization:** documentation of data (formats, content, etc.) and the metadata that describes it and makes it discoverable and re-usable.

Different levels of re-usability can be characterized by a lesser or greater reliance on these additional pieces of information. A discipline-specific policy discussion will be required in order to decide an appropriate level of preservation and re-use. 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 these requirements. However, we suggest a series of pilot projects that could supply a set of building blocks for the construction of this system. Finally, there is an exploration of which cultural norms would need to be shifted for broad acceptance of open access to data as the new paradigm. We

suggest that these issues not be overlooked; willing compliance of the research community is critical for the success of the open access movement. Finally, we make some suggestions for ways that the NSF can accelerate the evolution of the open data landscape. In addition to funding various pilot projects that can provide critical pieces of the open access infrastructure, the NSF could improve communication and set ground rules via the following recommendation:

- NSF should highlight and disseminate best practices for data archiving and sharing. This could take the form of publishing 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 those researchers seeking to emulate these exemplars.
- NSF should develop guidelines for trusted repositories. NSF does not have the infrastructure to store all of the research data that scientists may want to archive. Researchers will use a variety of repositories of varying quality and sophistication. Some minimum set of requirements should be established so that scientists know that they have done due diligence by storing 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 Open Data Policy

For informational purposes, the new NSF Open Data policy presented in NSF 15-52 is summarized here for those researchers who may not be familiar with its contents.

From NSF 15-52:

> "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" of any journal publication or juried conference submission into 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. The first workshop, held in November 2015, 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 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 discussions, 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. 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 towards the open sharing of data is the only way to accomplish broad adoption of the intended goals. 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 other 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[5], 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.

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.

Levels of Complexity in Data Curation and Sharing

A critical consideration in any discussion of data sharing is deciding what kind of data are to be shared[6]. This establishes the expectations for re-use and sets forth the requirements on what associated information must also be preserved with the data to make it accessible and understandable to the target consumers. An outline of the different levels of usefulness for data preservation and sharing can be found in the following diagram. Here, each higher level of usefulness for a dataset 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".

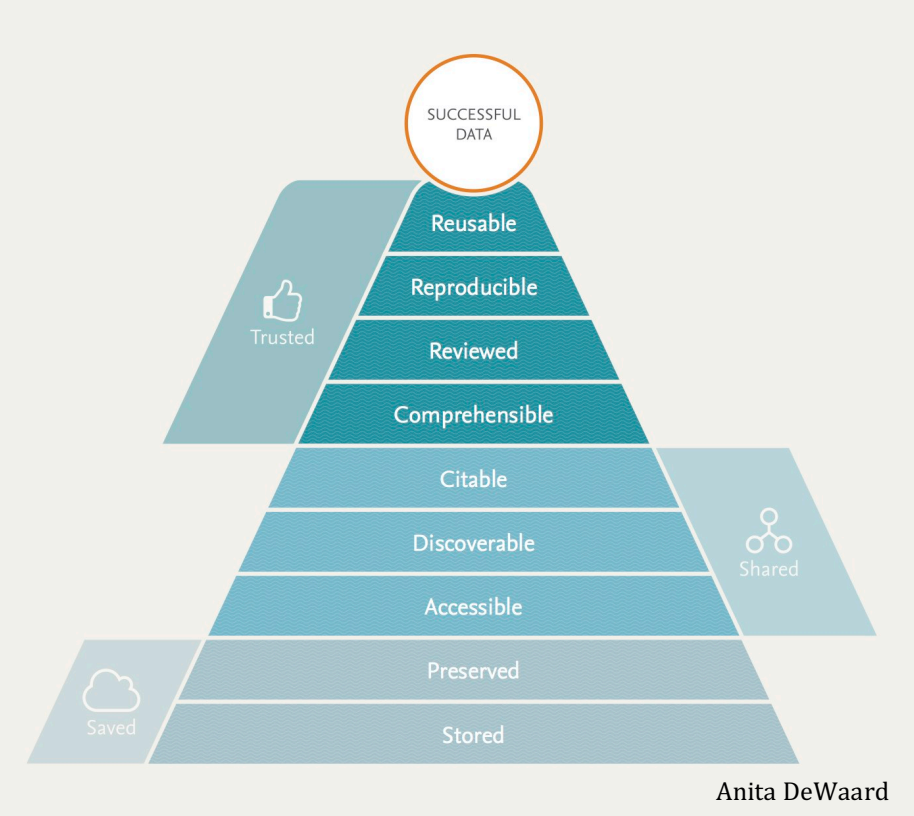
Beginning at the bottom, at the most basic level:

---

[5] https://www.force11.org/group/joint-declaration-data-citation-principles-final

[6] 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

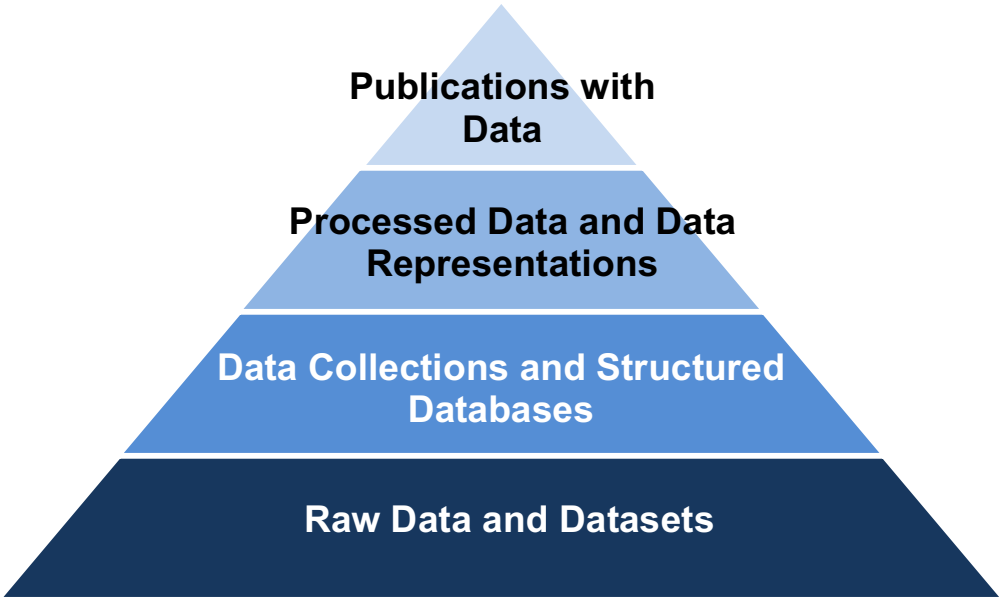
- "Stored" data exist in some form, somewhere. It is not necessarily archived, nor is the storage medium guaranteed to exist in the future.
- "Preserved" data are stored in some sort of archival format and placed in a location such that they will be available some time in the future.
- "Accessible" data are stored in a location that can be accessed online.
- "Discoverable" data are indexed in some manner allowing them to be found in online searches for relevant results.
- "Citable" data have an identifying marker, such as a DOI, that allow them to be referenced, and, more importantly, allow their impact to be measured.
- "Comprehensible" data carry with them accompanying descriptions of their content and the methods used to derive them.
- "Reviewed" data are curated, with 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., to allow the data to be processed or analyzed in a different manner than that which produced the original data/results.



As these descriptions show, making data increasingly useful for others when shared requires an increasing investment in information, processing, and curation.

Therefore, when setting goals for the 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. 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 domain and on whether or not the data is intended for re-use or re-analysis. Sharing the numerical results backing the figures of a given publication allows comparison and potentially 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-analysis. 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. These considerations are key ingredients in forming any decision as to what information should be shared.



There is ample evidence that the conscientious calibration, curation, and preservation of research data has immense benefits. For example, the data in the Hubble Space Telescope archive 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 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 a 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 the same impact in terms of citations as papers from the principal investigators.

Further evidence comes from the Sloan Digital Sky Survey, 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, many of which were written by scientists having no affiliation with the Sloan project. SDSS data have also been widely analyzed by citizen scientists, a number of whom have become co-authors on professional research publications.

And most recently, a team of ~1,000 researchers shared data, simulations, and processing algorithms in the discovery of gravitational waves with the Laser Interferometry Gravitational wave Observatory, LIGO, 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.

Steps Toward Open Access

As the previous sections have indicated, preparing data for sharing requires effort, but there can be tangible benefits. One exercise of the workshop was to propose concrete steps that may be implemented in order to move the MPS community toward open access. A general consensus arose around the 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 open access:

Data upon which publications are based should be available in machine-readable digital format, and persistently linked to those 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. 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, and, from an implementation perspective, would be a simple extension to most data-management plans.

Clearly, this would be a first step along a path towards 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 was that we are not ready for open access to research data if we wish to base this on existing infrastructure. To advance toward the goal of universal open access, new capabilities for data/knowledge archiving systems will 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 in a re-usable manner, beyond just the data:

- **Software:** the software used to create, process, and analyze the data
- **Workflow:** instructions, frameworks, or scripts use to run the software
- **Software environment:** a specification or a instantiation of the requisite operating system, architecture, libraries, etc., that are necessary to run the software/workflows
- **Simulation capabilities:** the capability to run the software with different parameters than used to generate the original data
- **Documentation:** a description of the software, workflows, and other information describing how the data were derived, processed, and analyzed.
- **Data characterization:** documentation of data (formats, content, etc.) and the metadata that describes it and makes it discoverable.

Obviously, the depth to which these elements need to be included will depend on the ambition of the open-access policy and the nature of the research. If the aim is less than full re-use, then the requirements will be less stringent.

Technical/Infrastructure Requirements and Capabilities

The components necessary to enable open access to data are distinct from the required infrastructure. 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 curated data. To this end, a number of technical attributes/requirements and capabilities were discussed for open-access infrastructure. The capabilities listed below were considered important/essential in any universal infrastructure for open access to data.

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

- **Federated storage infrastructure:** data will be stored in a variety of archives; these archives should be globally accessible and interoperable
- **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
- **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 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

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 may well be discipline-dependent, adding to the complexity of the task. A further obstacle arises in this respect because the benefits resulting from open access to research data are difficult to quantify broadly. This is partly because the quantity of data that has been open for public access has been minimal in some fields, but also because the cost model for calculating the benefits of additional 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 an architecture that allows preservation and sharing would be straightforward. 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 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. Even with this procedure, 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.

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? [7]
    - 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 might include citations of data usage, access records, etc.
- **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.

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. We list them here as potential ingredients for a successful open-access culture. Clearly, targeted infrastructure investment might be of some benefit in hastening the establishment process and for community-building.

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 head. As mentioned in the previous section a globally discoverable and searchable data

[7] 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.

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[8];
- [others??]

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 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. 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.

Pilot projects

Finally, we 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

[8] https://www.computer.org/csdl/mags/cs/2012/04/mcs2012040026-abs.html

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

- Attach additional funding to grants, or have separate RFPs that encourage different 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

- 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.

Suggested actions for the NSF

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:

- NSF should make a point to highlight and disseminate best practice for data archiving and sharing. This could take the form of publishing excellent 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.

- NSF should develop guidelines for trusted repositories. NSF does not have the infrastructure to store all of the research data that scientists may want to archive. Researchers will use a variety of repositories of varying quality and sophistication. Some minimum set of requirements should be established so that scientists know that they have done due diligence by storing 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.

Outreach to MPS Fields and Community Discussion of this report

This report is intended to represent the consensus opinions of the broader MPS community. It is a draft document, to be presented for discussion. It should be discussed for broader audience wherever and whenever an opportunity presents itself. We hope this will include such venues as

- Society meetings (APS, ACS, AMS, etc.), DPF, APS meetings
- AAAS
- Society publications/newletters
- OSPs, in order to involve the university research communities directly