National Science Foundation
WHERE DISCOVERIES BEGIN

Final Report:

An NSF-Wide Workshop to Explore the Prospects for a Common Response to the Requirements for Public Access to Research Data[1]

[1] Funded by NSF-PHY-1654844

TABLE OF CONTENTS

1. Introduction

This report reflects the deliberations and conclusions of an NSF-wide Workshop to explore the prospects for a common response to the requirements for public access to research data. Representatives of almost all of the NSF Directorates convened in Alexandria, VA, February 22 and 23, 2018, to review a diverse and multi-disciplinary collection of projects and workshops that have been conducted in the recent past. All of these activities were focused on aspects of public access to research data, from broad surveys of entire areas of research to development projects aimed at prototyping specific infrastructure for data access and discovery. Many of the PIs or co-PIs of the various projects attended the workshop or provided material for discussion.

The workshop had three major points of focus:

1. The sharing of lessons learned and findings as collected from each of the scientific communities
We expect that many of the concerns and suggestions will be similar across the different communities represented by the previous workshops. Yet, discipline-specific distinctions will almost certainly exist. What level of commonality can be found? And, is the common ground a sufficient basis for common guidelines?

2. Linking communities with common interests in knowledge preservation and access
Each scientific community has a group of experts or specific projects that focus on data (and software) preservation and access. Moving forward, how can these groups be leveraged to provide continued information and guidance as the scientific views of open access, the needs of the public, and scientific policies evolve over time? What fruitful interdisciplinary dialogue or coordination between these groups can be established?

3. The formation of suggested requirements related to knowledge preservation and open access
Given the results of the discussion on commonality and future collaborations, what inputs might be suggested to the NSF in the following areas:

- What might general NSF guidelines for knowledge preservation and public access look like, and how might they evolve?
- What information and recommendations should be given to reviewers and review panels in order that they can appropriately evaluate the knowledge preservation and sharing plans for submitted proposals?
- What guidelines or templates should be given to individual PIs as they prepare their proposals so that they are able to succinctly describe their compliance with requirements for preservation and access?
- How should the NSF best motivate and communicate these guidelines?

As the workshop progressed, it became clear that there exists **consensus** on many of the conclusions related to public access to data across the many disciplines represented by the NSF. In particular, all of the scientific communities represented recognize the importance of data sharing, moving towards FAIR principles (that data should be Findable, Accessible, Interoperable, and Reusable)[2] for research data. The participants felt that this consensus should **empower** the NSF to take a leadership role in all aspects of public access to data, from establishing policies to technical innovation to supporting workforce development, since it is clear that its scientific stakeholders support these efforts. The priorities for the scientific communities and the suggested areas where the NSF can have the most significant impact are presented below in light of this overall goal of NSF leadership.

2. Overview

Discussions at the workshop, based on the reviews of the specific projects as well as broader inference from the collective experience of the participants, revealed uniform support for the broad preservation and sharing of research data. More extensive sharing and re-use of data among scientists will accelerate the learning cycle, enabling access and analysis of the diverse data sets whose combination is necessary to address many of the complex problems facing today's society. Sharing these data with the public at large enables citizen science, promotes social and economic development based on research outputs, demonstrates good stewardship of publicly-funded research, and enhances public confidence in the scientific process by allowing results to be openly reproduced. All of these benefits are directly aligned with the mission of the NSF.

This enthusiasm was tempered by the acknowledgement that the scientific community currently sits at the very beginning of a long arc moving toward an eventual goal of public access and reuse of research data, expressed most clearly by the FAIR principles. That eventual goal can only be achieved by a series of targeted near-term steps towards these future goals. Many aspects of the scientific enterprise, including infrastructure, training, and scientific culture will need modification, amplification, and certainly innovation, in order to make significant progress.

These near-term steps should be driven by a clear set of policy guidelines around public access that are accepted by the research communities. As a leader in this domain, the NSF can monitor and enforce adherence by making future funding contingent on compliance. On a more positive note, the NSF can drive innovation and progress in these areas by targeted applications of funds that support the overall goals of the agency, including the current "Big Ideas" initiatives and the NSCI.

The current explosion of research data, the rising ubiquity of computation in research, the growing interdisciplinary research challenges, and the bewildering diversity of efforts aimed at

[2] M. D. Wilkinson, et al., "The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data **3**, 160018 (2016). doi:10.1038/sdata.2016.18. Information available at https://www.force11.org/group/fairgroup/fairprinciples.

imposing local order make for a situation ripe for agency leadership. The systemic issues that exist can only be overcome by leadership and coordination at a national or international level. The NSF can take the lead in establishing guidelines, standards, infrastructure, and sustainable practices by maintaining its focus on innovation while taking the long view necessary to develop a data-sharing ecosystem that will the basis for future research and discovery.

3. Data and Data-Sharing: Definitions, Modalities, and Other Considerations

3.1 What are data?

For the purposes of this report, we consider "data" to be the output of the research enterprise, however that is defined for each project. Since, as discussed, the eventual goal is sharing and reuse of scientific results according to the FAIR principles, "data" is the scientific compendium that encompasses a given element of research. In order to enable reuse, sufficient annotation and supporting documentation must be included with the basic scientific "facts" so that the facts can be understood by others. In this case, "facts" could be a series of measurements represented in numerical form, but they could also include audio and video recordings, text, physical samples, etc. Sufficient documentation could be as simple as clear and unambiguous column headings on a spreadsheet, but could include analysis software, simulation software, records of correspondence, published and/or unpublished descriptions, etc. The key word is "sufficient" to enable reuse.

3.2 What is sharing?

Continuing to take FAIR principles as a guide for this discussion, the "data" as described above should be Findable and Accessible. This implies that they are somehow indexed in an open location on the internet, and that they can be downloaded or used by interested parties. We will make a distinction here between "public" and "open" data. "Open" data implies, at least for our purposes, free access, whereas "public" could entail the imposition of various restrictions, such as subscription fees.

3.3 The "Public" as a Stakeholder in "Public Access"

As mentioned in the overview, sharing data with the public enables a dialogue between scientists and their de-facto supporters. Because federally-funded research is directly supported by a fraction of the tax income paid by ordinary citizens, those same citizens have a right to know what results came from their contributions. Allowing the public to access research results is a direct demonstration of the good stewardship of publicly-funded research. This argument is especially compelling given the billions of public dollars spent on research and development in the United States. At the very least, making the public aware of research by allowing access emphasizes the new knowledge and the process of discovery, from which the public derives immense benefit. We, as scientists, cannot necessarily anticipate all of the ways our results could be used by enterprising citizens or other scientists. The act of sharing enables whatever

added value might be derived from our results. Sharing research data with the public, particularly when coupled with educational initiatives, can democratize access to science, equalizing opportunity for scientific exploration by anyone. Providing these data opens opportunities for all to explore and to be engaged by science, potentially increasing the diversity and inclusivity of the scientific workforce. This engagement enhances public confidence in the scientific process by allowing results to be openly reproduced.

4. Public Access Projects: Results, Consensus, and Issues

As mentioned in the introduction to this report, this workshop revealed a broad consensus across the scientific communities supported by the NSF that data sharing and reuse can accelerate the scientific process, and that data sharing should extend to the public sphere to enhance public engagement in and support of scientific research. However, there is also a host of problems and issues to overcome to make this vision into a reality. Here, we summarize the scope of the various projects whose conclusions, taken together, lead to these consensus statements, and then outline the issues that must be solved.

4.1 Public Access Projects

The workshop reviewed the results of a number of different projects that were funded by NSF to address questions in and around public access to research data. (See Appendix A for a list of projects.) The projects themselves were quite diverse, encompassing workshops, tool-building and exploration, and innovative infrastructure support. Taken as a whole, however, the projects addressed almost every aspect of problems and issues surrounding public access to research results. For example, Attewell (Award 1243785), Esteva (Award 1555458), Hovy (Award 1450545), Lehnert (Award 1449298), Mayernick (Award 1449668), O'Grady (Award 1449499), and Webster (Award 1451374) examined various aspects of indexing, discovering, accessing, and analyzing shared data. Esteva, Mayernick, O'Grady, and Webster included specific recommendations and focus on the need for standards around data and access. Hildreth (Award 1457413), O'Grady (Award 1450894), and Ruggles (Award 1451112) presented the views of their individual research communities on many aspects of public access. Hildreth, Lehnert, and Webster included a specific focus on tools for researchers. Ruggles along with Berman, Lehnert, and Stodden (Awards 1649545, 1649703, and 1649555) specifically addressed aspects of Data Management Plans.

4.2 Statement of Consensus

As the projects were discussed, it became increasingly clear that the problems encountered and the recommendations made in the various projects and reports were **nearly identical** across all of the disciplines represented. While this may not be surprising in hindsight, it is striking given the diversity of topics, methodologies, backgrounds, and personnel involved in research across the agency. This leads us to make the following statement:

There exists consensus on many of the conclusions related to public access to data across the many disciplines represented by the NSF.

In particular, all of the scientific communities represented recognized the importance of data sharing, both among scientists and with the public. This implies broad acceptance of the need to preserve data and the knowledge of how to interpret it. To make a concrete recommendation, the participants felt that setting the adoption of FAIR principles (that data should be Findable, Accessible, Interoperable, and Reusable) as a goal should be the driver for policy considerations moving forward. This would give a clear, yet ambitious target and a clear statement to researchers and policy makers as to the direction of data sharing and public access policy.

As a consequence of this unanimity, the participants felt that this consensus should **empower** the NSF to take a leadership role in all aspects of public access to data, from establishing policies to technical innovation to supporting workforce development, since it is clear that its scientific stakeholders speak with one voice in support these efforts. In contrast to other national funding agencies with diverse science bases, NSF representatives can speak from a position of unified strength on the issues discussed here as national policy is created and evolves.

4.3 Community Priorities and Issues

In arriving at the consensus statement, above, many issues were discussed. This section presents an overview of those that emerged as community priorities, and discusses some of the more pressing problems in this domain.

First and foremost, there is a need for a clear policy statement and guidelines directed at the preservation and sharing of research data. The statement that the adoption of FAIR principles is the goal of the agency would be one way of accomplishing this.

Any discussion of consensus should also address a common set of problems envisioned with the construction of a publicly-accessible data-sharing ecosystem. As was the case with the expected benefits of data-sharing, current issues and anticipated problems were raised by many of the projects. 1) One set of issues is related to the technical aspects of data sharing. For example, many disciplines lack standards for data formats or metadata descriptions, making sharing and interoperability difficult. Individual investigators often lack tools that allow them to prepare easily their data for preservation and sharing. Typically, individual archives are not connected to each other in any useful manner, impeding data discovery. A distributed computational infrastructure that can access diverse datasets from different repositories to produce aggregate analyses does not exist. 2) A second set of problems revolves around the impact of public-access mandates on the individual investigator. Direct conflicts can exist between the need to share data and other research products and the intellectual property of the researcher. This leads to questions of embargo policies, licensing, and proper citation credits. Issues of data misuse and liability also arise in this context. Current uncertainties and potential policy variations across funding agencies drive investigator angst. Policy flexibility is also an

important issue. Often, the information required to make a dataset reusable varies at the individual project level. The researcher must have the primary role in deciding how to meet policy guidelines. 3) A third set of issues relates to costs. Sharing data levies a cost to the researcher, in time, infrastructure, and effort, and includes costs to archives, as well as to funding agencies. Determining the appropriate costs and their distribution is a critical and likely contentious set of decision points. 4) The fourth and final set of concerns are cultural, and thus harder to address. Activities around data sharing do not currently receive a sufficient level of credit in scientific circles to incentivize broader participation or acceptance. Several issues are at play here, including data citation practices, promotion and tenure requirements, the effort cost for data sharing, and the lack of clear policies.

None of these problems are insurmountable. With clear policy statements, coordination, and seed investments in innovative infrastructure, the NSF is positioned to be an international leader in creating the new research data ecosystem.

5. Roles for the NSF

By virtue of its position as a funding agency, the NSF has a variety of means for influencing the scientific community, and, through coordination with other agencies and international entities, influencing national and international policy. With decisive action and policies, the NSF can take leadership roles in all of the areas required for moving the scientific community along the arc towards a shared-data research paradigm. Specific areas of focus could be:

- Formulation of requirements
- Sustainability
- Capacity building, including in the areas of:
    - Workforce development/training
    - Cultivating curation activities, tools
    - Infrastructure
- Cross-disciplinary dialogue and coordination
    - Around infrastructure
    - Around policies and standards
- Driving culture change

Each of these potential areas of influence and development is explored briefly in separate sections, below.

Broadly speaking, the NSF has many different ways to influence the adoption of public access to research data as the norm - much as Broader Impacts are now the norm for all grants. For example, the strongest possible policy action would be for NSF to change the funding model to require a change in behavior from all scientists, such as withholding some fraction of grant funds until data were deposited in a certified repository. This policy action would certainly change the operating procedures of researchers; however, it may not be suitable at present for those directorates that currently do not have well-developed public archives. A strong but more

subtle change might include the path initially taken with Broader Impacts that provides a reward for *explicit and substantial data sharing* in the Scientific Merit and Broader Impact sections of grant proposals, along with the action of following-up during the review of a grant Final Report on whether Data Management Plans have been followed. At the very least, a continual conversation with stakeholders to define and drive policy changes is required. However, interminable dialogue is not a policy solution.

From a funding standpoint, NSF could make a number of choices to motivate the development of the infrastructure necessary for public access. For example, NSF could fund research that studies the process of sharing itself, looking at the impacts on researchers of the need to prepare their data for sharing, and what motivates researchers to share more broadly. Identifying motivators is a powerful step to understanding how to motivate data sharing. Research could also be sponsored to examine the impact of reuse on scientific productivity, public perception, public engagement, as well as how to increase the influence of all of these. Specifically directing funding to projects that reuse data will spur innovation in reuse and promote the reuse of data. More broadly, meta-studies could be conducted to discern which funding mechanisms are most likely to result in greater adoption of data sharing. Some of these ideas have recently appeared in a Dear Colleague Letter (NSF-18060), which we take as an encouraging sign.

Effecting a cultural shift in the way research is conducted and shared, and developing the infrastructure to support the public sharing of data, while likely beyond the scale and scope of the NSF alone, is an important goal towards which the NSF should lead. The NSF clearly has a role to play in funding and promoting the adoption of innovative solutions, and then forming larger partnerships with other funding agencies, disciplinary organizations, research institutions and individual researchers to build capacity.

It's clear that investments in these areas align well with the stated goals of the NSF in the area of data science and computational infrastructure. Data sharing will be integral to Harnessing the Data Revolution (HDR), for example. The training and workforce development discussed here is squarely within the education scope of that "Big Idea." [3] In addition, the cyberinfrastructure required, in addition to being relevant for HDR, can also be considered as relevant under the NSF contributions to the strategic objectives of the National Strategic Computing Initiative (NSCI)[4]. Thus, this discussion is not at the fringes of NSF's scientific and public mission. Rather it relates directly to many things close to the core of NSF's scientific identity.

6. Formulation and Implementation of Requirements

A recurrent theme during the workshop discussions was the need for clear and enforced guidelines for public access to research results. If NSF wishes to create a culture of sharing results between researchers, and, more broadly, with the public, guidelines mandating this

[3] https://www.nsf.gov/news/special_reports/big_ideas/harnessing.jsp
[4] https://nsf.gov/cise/nsci/

behavior need to be enforced through mechanisms directly related to funding. It is certain, however, that the technologies, attitudes, and infrastructures around public access will evolve rapidly over the next few years. Therefore, a continuous series of dialogues should be maintained, within and exterior to the NSF, in order to create a sustained series of conversations around appropriate guidelines and their potential modifications to align with future trends. The main instrument for guiding policies is the Data Management Plan (DMP) required of all NSF grant applications. An extended discussion on the role of the DMP and how it might be modified or augmented yielded the thoughts expressed below.

6.1 Data Management Plans

The DMP remains the main tool for guiding researchers through the process of preserving and sharing their research results. Enhancing the infrastructure for submitting, processing, and verifying DMPs would provide a relatively simple means for advancing the state-of-the-art in the sharing of research results. First and foremost, the workshop participants strongly felt that DMPs should be both **verifiable** and be **verified** at some point during the process of reporting on the progress of a grant. The obvious time to do this would be during the acceptance of the final report. Something as simple as adding a small number of questions to the reporting functions in research.gov ("Have you deposited your research results in an archive?", "If so, give DOI and access information", etc.) would signal a commitment from the NSF to the policies underlying data sharing. Clearly, program officers would have to flag non-compliance as an issue in accepting a report. Currently, they do not even have the information to do so. Achieving more robust and automated validation will certainly require modifications to the way DMPs are currently structured. Making them machine-readable, i.e., electronic form-based, would allow verification; more elaborate infrastructure is necessary for automatic validation that the processes laid out in the DMP have been followed. One could imagine, for example, procedures that check if a specific dataset is accessible in the target archive without any human intervention.

In order for the validation to make sense, clear guidelines are required. This point cannot be over-emphasized. If the goal is to proceed to sharing of results based on FAIR principles, more detailed guidelines should be developed, including some minimum standards underlying each of the expected attributes. These should be clearly communicated, potentially providing more resources for the average investigator. Sharing exemplars of best practices, models of procedures, and providing examples of appropriate repositories for the deposition of a given set of research results would give researchers a much clearer idea of what is expected and how to achieve that result.

This is an area where NSF can seize the forefront and lead by providing clear guidelines and expectations to researchers.

6.2 Community Engagement

As mentioned above, it is highly likely that the technology that enables public access to research results will evolve rapidly over the coming years. While research community perceptions and those of the public may not change quite as quickly, these developments suggest that an ongoing engagement between various aspects of the NSF advisory structures on these topics would be beneficial in providing feedback on potential changes in policy and opportunities for investment in research. In particular, the different advisory councils can play a leading role in this area. The Advisory Committee for Cyber Infrastructure (ACCI) could have a standing subcommittee for public access that would span the disciplines represented on this NSF-wide committee. The topic of public access could also be taken up as a recurring theme in the advisory councils of the individual discipline-focused advisory councils. Committees of Visitors could be instructed to pay particular attention to divisional support for public access and individual programmatic efforts in this domain. In a direct outreach to the research community, standing interdisciplinary workshops, such as the one summarized by this report, could explore different aspects, implications, and difficulties encountered in the implementation of public access to research results. (Certainly, the participants felt that this first interdisciplinary workshop was very successful in presenting a broad view of disciplinary needs, while establishing a baseline consensus position on many aspects of public access.) To address the "public" consumer side, it will be important to open avenues of communication between producers of research results, those who support their efforts, and those who might use them. Gatherings could be organized, for example, that include leadership in higher education, researchers, foundations that support them, societies like AAAS, and industrial partners. These discussions should result in a sharper focus on how research results are likely to be reused, and how best to share them.

7. Sustainability

As scientists, through collaboration with repositories and infrastructure, offer public access to data, important sustainability question arise. Given that making data accessible and understandable to the public is costly, it is not clear that all data in all stages of preparation should be made public. This question is particularly relevant for data that must undergo extensive cleaning and processing before it is useful. Likewise, although storage costs are falling, preserving data requires ongoing effort on the part of repositories, with associated costs. Decisions must be made about which data to preserve (e.g., some data, like historical specimens, are irreproducible, while other data types, like some DNA sequences, may be easily regenerated), how long to preserve them (given that we cannot know their future utility with certainty), and how to fund ongoing data preservation and access. Should current researchers be expected to fund the preservation of their data outputs indefinitely? Current policies and culture generally support the inclusion of data publication costs (e.g., to pay a repository that will house the data) in proposals, but those costs generally are not sufficient to support the repositories in the long term. Economic models that use the intrinsic value of the data to generate future income may help to sustain repositories, but may also be in conflict with the

public access goals of democratizing access to science and equalizing opportunity for scientific exploration.

This continues to be one of the most contentious issues in the realm of data preservation and sharing. The NSF should continue to show leadership in this area by working with investigators and other agencies to develop appropriate sustainability solutions for data storage and access. Without this, the future of public access to research data is in doubt.

8. Capacity building

A systematic approach is required to build the capacity necessary to make data sharing and public access the norm. In order to achieve this norm, the process of data sharing must be relatively straightforward and everyone needs to believe the process is beneficial/required as well as know how and where to archive and share their data. Below we discuss three areas of need: Workforce development/training; Investment in curation activities and platforms; and Infrastructure development.

8.1 Workforce development/training

For data sharing to become normative, training is required at all levels of the academic life cycle. However, for this to occur - training must be available and incentivized. NSF can encourage this development by partnering with scientific societies, industry, repositories, libraries and institutions to develop training capacity. Such capacity might include curricular development at colleges/universities and focused workshops at society meetings. NSF can incentivize the use of such training by making data education training a required component of Postdoctoral and Graduate student mentoring plans within individual grants, by enhancing the protocols for training requirements in training grants, and by rewarding such data education in individual GRFP applications.

The aim here is to achieve diffuse education throughout a field or fields - not only to individual awardees - and incentivizing this diffusion will be facilitated by supporting different kinds of research that focuses on building and sustaining capacity. For example, research on how training can most effectively be marshalled to lead to capacity building will be essential. This could entail funding calls through specific NSF directorates (e.g., SBE or EHR) as well as interagency programs. Alternatively, providing funding for sustaining efforts in capacity building (rather than just innovations in capacity building) will also be critical. In this latter case, the value-added could focus on innovations in how to disseminate training/capacity building.

8.2 Investment in Curation Activities, Platforms, and Tools

While workforce training is critical to public access, appropriate platforms, tools and curation capacity that will enable researchers first to prepare for preservation of their data and the knowledge necessary to interpret it, and second, to deposit it in a trusted archive are also pressing needs. Aside from a few notable discipline-specific exceptions, the average

researcher funded by NSF has at her or his disposal only rudimentary tools for the preservation of research results. Depending on the project designated for preservation, a researcher could desire to archive data, software, workflows, computational environments, and even physical objects. Metadata vocabularies to describe these are often lacking, impeding progress towards the ingestion of the results into a repository. Repositories, portals, or other research platforms may or may not be prepared to accept the full variety of data that a researcher wishes to deposit.

Exploring and then filling the gaps between the researcher's raw data and an appropriately-curated, reusable dataset is a critical need if the goal of FAIR data is to be met. NSF could work in this space by bringing together research tool developers, scientists, archivists/data scientists both within and across disciplines. These groups can then establish the needs of the various research communities, survey what tools exist, and identify common gaps that may be served by more general tools. NSF investment in collaborative and resource-efficient development of tools that can be used to satisfy various groups could then follow. Coordination across disciplines can help avoid the prevalent tendency for each research group to write their own tools in the belief that their particular problem is unique. One means of financing this investment might be to allow supplements to individual grants that would support the systematic curation of data (or partnerships with repositories).

8.3 Infrastructure for Data Sharing and Public Access

For data-sharing and public access to become normative, a far better infrastructure is needed in support of these activities than currently exists. Making research products available to the public requires a secure and robust cyberinfrastructure to support archival storage, the indexing and searching of the products, and the capability for access, download, and, potentially, computation. Widespread data sharing between researchers presents more extensive cyberinfrastructure needs beyond the issues associated with public access. "Active" data that is being analyzed, for example, must be connected with appropriate computation resources and could be transient, with all of the extra book-keeping requirements that this entails. A continuum of resource needs exists to support ongoing science while providing a robust basis for preservation and public access.

While there is a great deal of variation across directorates, disciplines and projects, in nearly all instances the current infrastructure for public data access is recognized as ad hoc and piecemeal, without clear means of or plans for integration across platforms[5] (in other words, nothing along the lines of NCBI exists for NSF-funded, non-genomic research data). We desperately need a strategic collaboration across directorates not only to vet individual archival projects but to form a plan for how to best build and then bridge between these functional repositories, and to sustain them, in order to best achieve the FAIR goals. A strategic plan for

[5] Of course, there are exceptions for various research domains who have self-organized and created infrastructure appropriate for given research communities. In the biomedical arena, NCBI (https://www.ncbi.nlm.nih.gov/) stands as an example of a fully-funded solution that serves a huge community of researchers.

infrastructure development focused on support for preservation, sharing, and public access could include a focus on those elements of cyberinfrastructure that directly support access, education, and outreach to the public for the purpose of science learning and dissemination of results.

This is another area of development that clearly aligns with the broader goals of NSF.

9. Cross-disciplinary/Agency Dialogue & Coordination

Interagency dialogue and consultation (including discussions with corresponding agencies in other countries) on the issues surrounding public access to research results need to continue, both at the management level and between individual researchers. The NSF should continue its current leadership in these interagency discussions. Two focus areas stand out as particularly important: infrastructure and standards. On the infrastructure front, there are several initiatives that could pave the way for interagency cooperation and joint infrastructure development. The first of these would be the creation of a "developers' forum" where those building storage and other public access infrastructure from all different disciplines could interact. Currently, it's difficult to get researchers to attend gatherings on public access topics; the developers that are building and using the infrastructure for NSF-related projects have no or little means of developing a community based on open source sharing of infrastructure design, strategies, software, and development[6]. Facilitating these interactions could break down the silos that currently exist in which each discipline operates without much outside communication. This could lead to joint infrastructure projects, sharing of best practices (and software, and software development) and, in general, more cooperation and less duplication. This may also aid in the development of common, open standards for interoperability and sharing of research products. Funding to incentivize the re-use of software could act as an accelerant, and would be a natural extension of the funding for data reuse found in NSF-18060.

Standards bodies and societies, although generally disciplinary by nature, provide another avenue for coordination of data sharing efforts across disciplines. These organizations can serve to disseminate information to practitioners in their discipline and represent those practitioners in cross-disciplinary efforts to develop general best practices, infrastructure, and standards to make data FAIR and accessible to the public. Because scientific data knows no national boundaries, international coordination on standards development, training, and culture change is essential, and this is another role that could be filled by standards bodies and societies. These same groups could take a lead in developing training materials that need to be tailored to specific disciplines, such as selecting the correct repositories and metadata standards, while general training in using data sharing web services, best practices for data citation and attribution, and the use of persistent identifiers may be more efficiently handled by non-disciplinary organizations specifically focused on data sharing.

[6] Again, examples from other communities do exist, such as the BioStars effort (https://www.biostars.org/).

10. Driving Culture Change

The move to public access and data sharing requires a full scale culture change in how science is conducted. The participation of multiple stakeholders is required to achieve such change, and the route to that change must be intentional, built through a consensus driven process and clearly mapped. Stakeholders include NSF and other funding agencies, as well as disciplinary societies, academic institutions and industry and of course, individual PIs. While no single stakeholder can induce this change in isolation, we have outlined above a number of ways in which the NSF can be a leader in driving the needed change that weaves the goals of public-access into the entire fabric of the scientific discipline, its expectations and norms.

Driving culture change, while a seemingly daunting task, is no more than a series of near term steps that each build toward a particular set of goals, beliefs and standards. Prior to undertaking any concrete steps, NSF should convene a gathering of all stakeholders to begin the process and develop a sense of community, consensus and shared understanding. The initial step has been taken in the conception of the required DMPs (Data Management Plans) at NSF - but moving from these to full culture change is a long journey.

The key places where NSF could exert influence to accelerate culture change are: 1) in the consideration of individual grants by incentivizing individual PIs to integrate data-sharing and public access themes at the very beginning of their projects. Such incentivizing requires not only a reorientation in PI thinking, but also mechanisms for making the additional work involved a net gain for the individual rather than a cost. That is, we need means of selecting grantees that reward and perpetuate the desired behaviors; 2) by directing research toward public access by partnering across NSF directorates to incentivize the development and sustenance of curation tools and platforms as well as the development and refurbishment of discipline specific infrastructure through specific funding calls. These incentives aim to increase the ease and ubiquity of means for sharing data by funding the development and persistence of the necessary infrastructure and the science on which it depends; 3) by encouraging a shift in training programs by partnering with societies, other funding agencies and institutions to ensure and enhance capacity building in the workforce; and 4) by building a road map (via a multi-stakeholder community driven strategic plan) toward this cultural shift.

11. Conclusions

This workshop was one of the first times that representatives of most of the NSF directorates gathered to review issues, initiatives, and problems around public access to research data. The surprising level of commonality let the participants to assert that NSF should be able to lead agency- and nationwide efforts to solve the problems around public access. In this report, we have tried to lay out areas where the NSF can show real leadership on the national and even international stage. We hope that the agency, reassured by this unanimity of purpose, will continue to invest and coordinate projects in this area.

Appendix A: Projects Reviewed at the Workshop

One focus of the workshop was to review the outcomes of a series of projects related to public access to data that were funded by the NSF over the past five or so years. All of these had either PI participation at the workshop, PI remote attendance via video conferencing, or a set of comments forwarded to the workshop participants for discussion.

Paul Attewell, NSF Award 1243785, "Building an Interdisciplinary Community to Prototype Computationally-Intensive Analysis of Large-Scale Educational Datasets"

Helen Berman, Kerstin Lehnert, Victoria Stodden, NSF Awards 1649545 (HMB), 1649703 (KL), and 1649555 (VCS), "EAGER: Collaborative Proposal: Supporting Public Access To Supplemental Scholarly Products Generated From Grant Funded Research"

Maria Esteva, NSF Award 1555458, "EAGER: Collaborative Research: Evaluating Identifier Services for the Life Cycle of Biological Data"

Michael Hildreth, NSF Award 1457413, "Workshop Series to Gauge Community Requirements for Public Access to NSF-Funded Research"

Eduard Hovy, NSF Award 1450545, "EAGER: A Method to Retrieve Non-Textual Data from Widespread Repositories"

Kerstin Lehnert, NSF Award 1449298, "Geoinformatics Facilities Support: Integrated Data Collections for the Earth & Ocean Sciences: The Marine Geoscience Data System and the Geoinformatics for Geochemistry Program"

Matthew Mayernik, NSF Award 1449668, "EAGER: Repository Cross-Linking for Open Archiving and Sharing of Scientific Data and Articles"

Richard O'Grady, NSF Award 1450894, "Proposal for a Workshop on Reducing Barriers for the Management, Integration, and Public Sharing of Large and Complex Data among Biologists Working at Genome-Phenome to Macrosystems Levels"

Richard O'Grady, NSF Award 1449499, "Issues Related to Changing Practices Around the Publication of Data"

Steven Ruggles, NSF Award 1451112, "Public Access to NSF-Funded Research Data for the Social, Behavioral, and Economic Sciences"

Michael Webster, NSF Award 1451374, "Meeting: Advancing Accessibility of Digital Media for Biological Research in the 21st Century"

Appendix B: Workshop Participants

Xiao-Feng	Xie	xie@wiomax.com	WIOMAX
Zunjing	Wang	wang@wiomax.com	WIOMAX
Felice	Levine	flevine@aera.net	AERA
Kerstin	Lehnert*	lehnert@ldeo.columbia.edu	Columbia University
Mike	Webster	msw244@cornell.edu	Cornell University
Catherine	Casserly	cathy@ccasserly.com	Hewlett Foundation
Ramona	Walls	rwalls@cyverse.org	CyVerse
Maggie	Gabanyi	maggie.gabanyi@rcsb.org	Rutgers University - Protein Data Ba
Susan	Antón	susan.anton@nyu.edu	New York University
Jeffrey	Spies	jeff@cos.io	Center for Open Science (COS)
Stephen	Ruggles	ruggles@umn.edu	University of Minnesota
Michael	Hildreth*	hildreth.2@nd.edu	University of Notre Dame

*Workshop co-chair

References

[1] M. D. Wilkinson, et al., “The FAIR Guiding Principles for scientific data management and stewardship,” Scientific Data **3**, 160018 (2016). doi:10.1038/sdata.2016.18. Information available at https://www.force11.org/group/fairgroup/fairprinciples.