View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Purdue E-Pubs

Purdue University

Purdue e-Pubs

Charleston Library Conference

Open Access, Public Access: Policies, Implementation, Developments, and the Future of U.S.-Published Research

Alicia Wise
Elsevier, a.wise@elsevier.com

Amy Friedlander
National Science Foundation, afriedla@nsf.gov

Howard Ratner
CHORUS, hratner@chorusaccess.org

Judy Ruttenberg
Association of Research Libraries, judy@arl.org

John Wilbanks
Sage Bionetworks

Follow this and additional works at: https://docs.lib.purdue.edu/charleston

 Part of the Library and Information Science Commons

An indexed, print copy of the Proceedings is also available for purchase at: http://www.thepress.purdue.edu/series/charleston.

You may also be interested in the new series, Charleston Insights in Library, Archival, and Information Sciences. Find out more at: http://www.thepress.purdue.edu/series/charleston-insights-library-archival-and-information-sciences.

Alicia Wise, Amy Friedlander, Howard Ratner, Judy Ruttenberg, and John Wilbanks, "Open Access, Public Access: Policies, Implementation, Developments, and the Future of U.S.-Published Research" (2013). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284315237

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

Open Access, Public Access: Policies, Implementation, Developments, and the Future of U.S.-Published Research

Alicia Wise, Director, Universal Access, Elsevier
Amy Friedlander, Staff Associate, National Science Foundation
Howard Ratner, Director of Development, CHORUS
Judy Ruttenberg, Program Director, Transforming Research Libraries, Association of Research Libraries
John Wilbanks, Chief Commons Officer, Sage Bionetworks

The following is a transcription of a live presentation at the 2013 Charleston Conference. Slides and video are available online at http://bit.ly/1eVZHPd.

Alicia Wise: International interest in the open access policies that will emerge here in the US, framed as public access policies, is high, and the outcomes of these discussions will impact the future of research and scholarship around the world. The devil is in the details. It can seem immediately obvious how to structure a public access policy, but to make one work in practice over the long term in ways that are affordable for all participants is quite a heady challenge.

Today, we have four highly influential people involved in these debates from different perspectives. First, we will have Amy Friedlander presenting the perspective of a funding body. She leads for the National Science Foundation on public access policy for full text. Then we will have Howard Ratner presenting a publisher's perspective, formerly from the Nature Publishing Group. He has now joined the CHORUS initiative as its development director. Third, we will have Judy Ruttenberg presenting a library perspective, representing ARL and the SHARE Project, and closing we have John Wilbanks who is presenting the researcher's perspective on these debates. He comes now from a small business, a start-up, which is heavily reliant on access to information, in particular, data, and I think he will have an interesting perspective for all of us other stakeholders in what we should be thinking about as we move forward. We are going to save questions for the end after the four speakers. We will have 10 minutes at that time, so a challenge for you: we are looking for succinct, insightful questions that will help us pull together the synergies or tease out the differences between these different perspectives, and I will be succinct and stop there. Amy, over to you.

Amy Friedlander[1]: Good morning. Well, it is a pleasure to be back in Charleston, and it is a pleasure to be seeing so many friends who I know through e-mail in three dimensions. There is really no substitution for real time. So let us get started—not enough time for reminiscences of Charleston and the 1970s when I was a student here.

So as you all know, on February 22 of this year the U.S. Office of Science and Technology Policy issued a memo that provided guidance for enhancing access to the results of federal investments, and its terms are well known to you. Fundamentally, it asked us to balance public/private partnerships, to do all of this within existing budgets, to make publications available within a guideline of 12 months, although that could be changed based on criteria that have yet to be determined. It asked us to have analytics, and it asked us to provide for unauthorized bulk download of online articles and a number of other features. Iam going to assume that you all memorized it and read it to yourselves every night. So where are we? Well, I am proud to report that the National Science Foundation, indeed, submitted its plan within the required 6 months. I can tell you it was roughly 2:05 in the afternoon of Thursday, the 22nd of August; not that I had anything to do with this. The plans will be made public after they are approved by OSTP

[1] Since this authorship contribution was done as part of official duties as a National Science Foundation (NSF) employee, the work is a work of the United States government and as such is in the public domain.

Copyright of this contribution remains in the name of the author(s).
http://dx.doi.org/10.5703/1288284315237

and OMB. There may be a set of exchanges in between what is considered final. They will be posted on the NSF web site as well as on the open government web site at that time, and I am sad to tell you I have no more insight than you do on the timeline. So whereas I had hoped when I accepted Alicia's gracious invitation this summer to be talking to you about what we plan to do, today, I will be talking to you about how we developed the plan to do it.

How did we develop it? We had four modalities, if you will: we wanted to collaborate, we wanted to listen, we wanted and do continue to want to leverage existing resources and capabilities in the many sectors that are engaged in making information available to the public, and we wanted to learn from prior experience. What did we do to collaborate? Well, it turns out that NSF had a bit of a head start on this. In spring 2012, you may know that our then-director Subra Suresh organized something that is now called the Global Research Council, and public or open access was one of the areas of study that came out of that. This provided us, then, an opportunity to begin looking foundation-wide at what would it mean to take the foundation into what we call a public access stance. It turned out, as many of you may know, that some of the directorates had already begun to engage in this activity. The Mathematics and Physical Sciences directorate, for example, was a leader in this area. I was then attached on detail to Education and Human Resources, and I came back to my home directorate of Social, Behavioral, and Economic Sciences to work with Myron Gutmann, the NSF Assistant Director for SBE, who was then leading the initiative.

We had a period of about 8 to 10 months in which [we needed] to understand a lot more about the agency and how public access would, in fact, affect every corner of the agency and to seek out our counterparts at other federal agencies that had, in fact, also begun to think about what this meant. When the memo finally appeared at noon on the 22nd, we were ready to begin to go, so to speak, and the group that we have formed with the other science agencies—NIST, NOAA, Department of Defense, Department of Energy, NIH, NASA—we were the nucleus then of what became an interagency working group. This informal interagency working group was split into two broad components. There was one on publications and there was one on data. There are tendencies, as you know, to see this issue as bifurcated into two tracts. We have publications over here and we have data over here, but we all know that they are intertwined, and we should be thinking about them as a continuum of products resulting from federal investment. I was the executive secretary for both of the interagency groups, and I had an interesting perspective on what was evolving. Within the foundation, a steering group was stood up to lead the development of the plan within the foundation. I was a member of that steering group, and then I chaired the subgroup on publications. In addition to that, we worked with our National Science Board which is the policy-setting body. As part of the listening enterprise, you may also know that we organized two public sessions for a period of four days in May, and then we talked to a lot of you. In fact, the first people to come visit me were SPARC, NCAR, and ARL, so I started with the library community. We had a lot of publishers who came to visit us. We continued to talk to our program officers, and we talked to our administrative staff again. The implementation is going to ride on top of the already heavy workload of the NSF staff, so early on it was apparent to us that we had to think of this not only in terms of its direct impacts on the immediate, if you will, consumers of what results with federal investment. We have to look closely about the operations, and, as we move into implementation now, very quietly, we are thinking more and more about the operational side of the house.

What were the issues that came up? Well, they are not going to surprise you. They are the usual ones. Was the repository going to be centralized or distributed? And, in fact, when we went through that conversation, it was not just the specifics of how do you build a distributed repository. It had to do with a lot of questions about data management, data storage, and who has custody. Once you claim a federal right in something, or claim federal ownership of

something—not that we are claiming that—well, there are certain responsibilities and expectations that are attached. Then, who is going to have access and access to what? This, for the publication side, frequently comes down to a not-so-simple discussion, Alicia, not-so-simple discussion of delay period or the period of the embargo, but, again, when you link publications to data, it ceases to become how many months but access to what; so let us take a few moments of my very short period and think about what happens when you publish an article in education remembering that the National Science Foundation funds fundamental work in education. Well, educational data, as many of you know, is bounded not just by human subjects, the common rule, not just by HIPPA, not just by the CFR, but also by legislation that specifically governs the management of educational data. So what do you do? You have a table that summarizes something, and we have said you have to make your data available and what happens? There are a lot of rules that restrict access to the underlying data, and they spill right over into publication. Other issues that came up have to do with compliance and metrics, what are the roles and, always, always, always, we have to be aware of change. I am afraid that I am going to bore you with that many times. In fact, those of you in the room who are mathematicians will recognize the images on these slides as fractals. They are from the Mandelbrot Set, and one of the significant properties of that set is that you can have a set of equations that end up in things that you cannot predict. So I chose that, obviously, intentionally, not just because we funded the research.

So what else can we leverage? Well, there are a lot of standards and best practices. I will not bore you by repeating a joke about the standards. We know that there is a lot of work that is going on in the Research Data Alliance that will bear on the way that we manage data and what we expect our investigators to do on behalf of data, and there are systems that we can look to. Now, I have been reproved for having said, in the past, that NSF does not have a repository. NSF, in fact, invests in many, many, many data repositories. However, we do not have anything analogous to the National Library of Medicine, to DTIC at the Department of Defense, to the National Agricultural Library, something that we support ourselves to support in-house researchers that could be used to manage the kind of item that we are talking about. So, of course, we started to look outside where could we borrow, where could we partner, how could we collaborate? We looked to the private sector for key pieces of infrastructure, notably CrossRef, FundRef, and ORCID, but there are others, and, again, I come back to this notion that we have to be prepared to change. When we look at these systems, when we think about it, we ask the question: is this extensible? How will it accommodate change?

What was the prior experience at NSF? As you may know, NSF is the premier, or we say we are the premier, civilian research agency that invests in all aspects of science and education. So we invest broadly in fundamental research in all aspects of science, not unclassified, and that means that our communities of researchers are highly heterogeneous. An economist really does not look like an engineer, and they do not necessarily look like a theoretical physicist. Their patterns of research are different, their expectations, the way they publish, the way they use information is different, and we need to be sensitive to that. We have been criticized, I know, because our data management plan seems to be quite abstract. It is abstract for a reason. It is so that many of the substantive decisions can be thrown back on the communities and exercised through the process of merit review.

Other things about us: our preliminary research indicates two important things about our investigators. One is that they publish in a very broad variety of journals. I asked to see the list of publication venues for our investigators just for FY 2011 and 2012, and I was handed a list with 55,000 entries in it. Since there are only 23,000 journal titles, I would say we have a disambiguation problem, but the point is made. The second thing we know is that our investigators are not exclusively NSF investigators. Most, if not all of them, have multiple sources of funding, and they have multiple sources of public funding, and this, in fact, is not unexpected. So if you go to our geosciences director, well, guess what? There is cofunding from

NASA. Right? Big surprise. If you go to our biological directorate, there is a lot of investment in NIH. NIH also shows up in engineering, things like robotics. NIH figures broadly in my home directorate of social behavioral, and so on; I will not bore you with the details.

Other things that we can leverage from our background is our data management plan. The fact that we allow data to be reported at the same level of granularity as evidence in our bio sketches, that datasets can be reported in annual and final reports so we are trying to maintain parity between data and publications, and that we can accept article processing charges as a direct expense on an application in a budget proposal which means, should someone want to go gold, we have a way of helping that move forward, should that be a decision on the part of the investigator.

Our approach, we hope, when we finally can share it with you is open, flexible, and incremental. We expect to continue to communicate with you. We want to minimize burden on program officers' administrative staff. That means we align whenever possible with what we are already doing, and we need to have high-level coherence. The last thing we want is an investigator who does something that is funded by multiple agencies to have to respond to different sets of rules even though the agencies are also at the same time respectful of their own background. So in my last 15 seconds, I hope you will agree with me that this is a good thing. We hope, in the end—it may be messy before we get there—to broaden access to research. We hope and we believe that information can be used to advance the foundation's mission and that this will provide a platform for innovation, and I think I am on time.

Howard Ratner: Thank you for giving me the opportunity to tell you about CHORUS. We have been very busy over the last few months. We delivered a proof of concept at the end of August. We have then incorporated as the not-for-profit CHOR, that is CHOR, Inc., and launched our pilot services at the beginning of October, and I also want to thank Amy for reminding me it is all about collaborating, listening, leveraging existing infrastructure, and also learning from prior experience, and I have decided that I am actually a serial not-for-profit creator, being there at the beginnings of CrossRef, ORCID, CLOCKSS, etc.; so there you go. Let us move on.

So as the first service of CHOR, Inc., CHORUS, or the Clearinghouse for the Open Research of the United States, now offers an open technology platform to meet the public access needs of funding agencies, researchers, institutions, and the public. Now, we are all about identification, discovery, access, preservation, and, finally—and this should not be forgotten—compliance. We want to meet the needs of all of our stakeholders, and it is important to note that there is no significant cost for the agency to use or participate in CHORUS. And why is that? Well, it is because CHORUS builds on the existing infrastructure. As you heard from John [Vaughn] before and Amy, FundRef, CrossRef, Prospect, CLOCKSS, Portico, and ORCID are all things that I have a little bit of knowledge.

Whenever you start a product or service, you need to identify the key stakeholders and figure out what might drive them to use such a service; or in other words, what do they want? CHORUS has identified around ten personas, but today I will only mention the top five, and I am only going to touch on the very top key drivers. You will be able to download the presentation and actually read through these more and, by the way, I am happy to accept feedback about any of these, so this is a collaboration because I want to make these personas as accurate as possible because it helps drive the product development. Keep in mind that personas are about what the stakeholders want and what they desire. It does not mean that every desire can or should be delivered. So here is Alan (referring to slide). He is arguably one of our more important personas. He is the agency department head, and he wants to meet the OSTP guidelines. He wants to measure grantee and agency compliance with those guidelines, and he also wants to show how the agency's investments are having impact or return on investment, and, obviously, he wants to provide access to his constituents.

Next, we have got Rachel the Researcher. Now, she wants to obtain funding for her research. She

wants to comply—and note here—with the funding agency requirements, not the OSTP, necessarily. She wants to know the sources of funding in her area of research, and she also wants to have access to the best available version of content in her research area. By best available version, if you read the footnote at the bottom there, it means either the accepted author manuscript or the version of record.

So we move on to Lottie the Librarian, which some of you in this room might be familiar with. She wants to have access to the best available version, too, but she is doing it on behalf of her patrons or for her own research. She also might want to do text and data mining for those articles all about things for her patrons. She wants to know that articles that are reporting on funded research are going to be readily available in perpetuity. She wants to help the researchers comply with the funding agency requirements, and we have heard that as a theme over the last two days. She also wants to build discovery tools for those very same researchers.

Now, Peter the Public, he has got some different drivers. He also wants to have access to the best available version, but he wants to do it because he wants to maybe research a problem that he has or he wants to drive some economic development that he has working on. He also just wants to see, generally, what the government is funding. He wants to learn about the impact of specific agency grants. He wants to understand the latest developments in science and putting things in context. He wants to have content connected to learning tools.

My last persona for today is Penny the Publisher. And I will go quickly here, but she wants to help her authors and her institutions comply with funder mandates, but she also wants to retain traffic on her journal web sites to better demonstrate value to librarians and also to attract whatever little left there is of advertising revenue.

So, how does CHORUS work? Remember this is working now. When a researcher submits a paper to a journal, they will interact with the management tracking system—and you can see some of the management tracking systems there—and they are all in various different states at this point. The system will then prompt the author to identify the funding agencies behind that article using a controlled hierarchical vocabulary as well as the grant IDs. Once the researcher puts that in, that is it. That is all that is required of the researcher—they are done. As far as they are concerned, they are going to let CHORUS take over and let the publishers take over after this. The paper will then go through the regular peer-review process and is published.

There is always this bit about preservation, and it is part of the OSTP memo, so when the paper is published, it is automatically deposited in at least one dark archive which includes CLOCKSS, Portico, or potentially another archival repository chosen by a funding agency. PubMed Central could be an example of that. The paper is then permanently archived in these repositories, but in most cases it is not made available.

How does the access part work? Well, articles are then made publicly accessible by the publishers host system in one of two ways. It is either after the funding agencies embargo period expires or is made immediately available if an author or funder has paid an article processing charge. They will get access no matter what to the best available version, again, either the accepted author manuscript or the version of record.

This is the discovery part of CHORUS. These articles, and this is important, can be discovered by a user using their favorite search engine. Now that could be the PAGES system that you see on the lower right side—that is from the PAGES system for the Department of Energy. We already are working right now with the Department of Energy. They have ingested 4,000 of our records, I am happy to say. It could be the CHORUS search engine. We have that going now—and that is live today—or it could be Google or it could be an institutional search engine, whatever it might be. All of this is fed by the CHORUS application programming interface, that API thing at the top. But this API also conserves text mining. It is also one of the things that agencies might want to do: to be able to grab this information, index it, process it for the purposes of discovery. That is also available via this

CHORUS API, and we are using the CrossRef Prospect Service to enable that.

Now, I mentioned that an important part of CHORUS is compliance, and this is very true. The CHORUS API feeds the CHORUS dashboard, which is now live in pilot, and then a government institutional publisher reporting system can use that to make reports. But also one of the things that CHORUS has done to demonstrate this, but also we found it to be very useful for some of the smaller agencies, is we have developed this dashboard. This dashboard actually reports on what the publishers are doing. So are they actually doing what they said that they are going to do? Can we actually identify the articles? Are the articles actually preserved? Are they publicly accessible, and, ultimately, do the articles have some kind of proper acceptable reuse license?

During the last few weeks I have had many questions about the CHORUS timeline. Here is a graphic showing our current plans (Figure 1), and you can see that we are now in our pilot phase, and we are also in our fundraising phase, and I have got a minute left here. So we will review our pilot phase in the new year which will then lead us right into our production phase in early 2014.

SHARE was mentioned before, so I wanted to say that CHORUS and SHARE did indeed meet in August 2013 to discuss our initiatives and explore areas of possible collaboration. So we agreed, especially at that time, to work jointly on persistent identifiers and metrics, and I am hoping, and John and I were just talking afterwards, that we will follow up and we will continue this conversation because ultimately CHORUS does not want to have anybody have any duplication of effort. These are our live services, and you can see that we have our live dashboard services, and you can access them on your phone or your laptops or whatever you would like. So we have got three live dashboard services with the USDOE, USDA. and NSF, okay, because this is, again, reporting on the data, and then we have our search service that you can all access. The information for all of this is also available at chorusaccess.org, and then we have seven pilot publishers. We have 13,000 public records that are in there, and we have 80 publisher's signatories.

Judy Ruttenberg: Hello. So, first, I want to thank Alicia for inviting me on this panel on open and public access and the OSTP directive and for framing this panel as a question of perspective. So I was invited on this panel to represent the librarian perspective, which is fair. I am a program director at the Association of Research Libraries and involved in the work that we are doing, this framework that we proposed with our partners in AAU and APLU called SHARE, and from that perspective, which I am honored to represent, the directive itself was an enormous achievement, one that ARL literally applauded when it was announced on February 22, as it happened, as our Board of Directors was meeting in our offices in DC. It was a victory and a long fought battle for

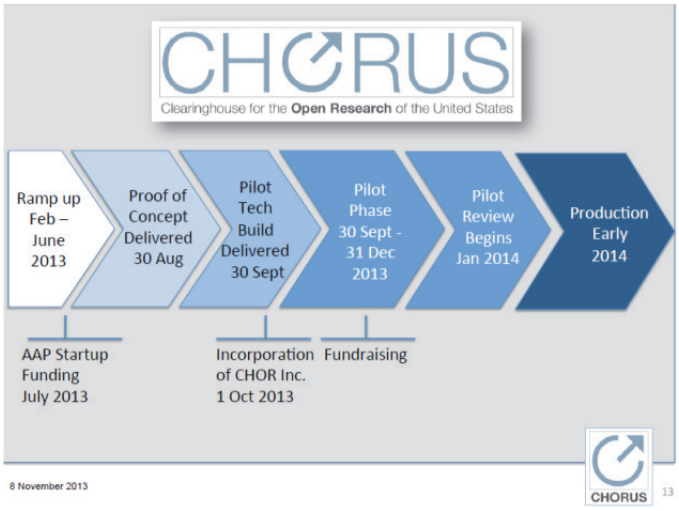

Figure 1. Timeline

public access to taxpayer-funded research and a policy that higher education is eager to see succeed as it is in alignment with the mission of research universities to create, disseminate, and preserve research, to see its research output used and reused in every sector to benefit society. And I very much appreciate the position and shared mission of society publishers expressed in yesterday's plenary session, I am thrilled to see major commercial publishers embrace the OSTP directive and eager to see those discussions continue.

So here is ARL applauding the directive the day it was announced, and our past president explaining that the memorandum reflects how twenty-first-century science is conducted, recognizing that it is data intensive, biased towards open; our Executive Director pointing out the years of investment by libraries; and the Academy to get to this point where our investments can be leveraged to see the success of this encompassing directive in opportunity. So I am not here, it is important to say, to pitch a product or to describe to you something working and functioning today. I am here to describe the opportunity that this directive gives us in libraries and higher education to do things differently, to skate to where the puck is.

After the applause subsided, we did get to work and issued joint statements with AAU and APLU at the May meetings that Amy described in support of public access publications and data. We produced a development draft of SHARE in early June for which we received invaluable feedback, including from colleagues around the globe. We formed a steering group in August, hired a consultant, Greg Tananbaum, in September, and secured additional Sloan Foundation funding in October. In the next week or so, we anticipate that the working groups just now being populated will begin their work building out SHARE. Those working groups include one for technology and standards, PI workflow, repositories, and communications. They will, of course, represent major players in the library community, but they will be broad-based and represent diverse constituencies. We have participation from leaders in the repository movement from commercial sectors including Microsoft, researcher driven products such as Mendeley, so the working groups themselves will be diverse and collaborative.

This is a chart from our June 7 SHARE development draft (Figure 2). This was sort of the

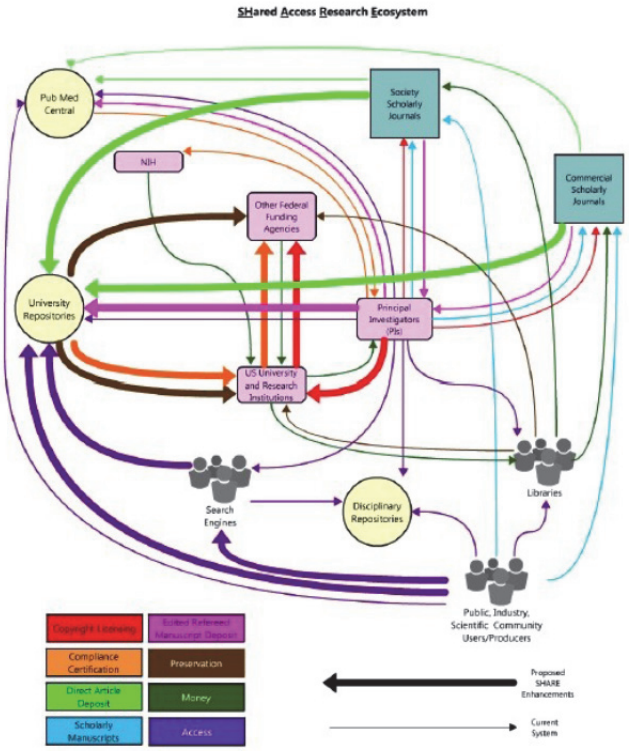

Figure 2. Shared Access Research Ecosystem

first thing that came out about SHARE. Uh huh, and it gets that reaction. It is meant to represent that ecosystem, right? The publishers, the funders, everybody who is sort of playing in this area that has to kind of—the devil in the details thing. But that box in the middle represents the PI, the researcher, who is at the center of this research process and directly accountable for the ways that the agencies will operationalize this mandate as well as accountable for any institutional or other funding mandate around deposit.

So this chart is overly complex. Some have suggested messy, confusing, even, and that is the world we live in and the world that SHARE is aiming to help simplify. Okay? But I think it is important, so just bear with me to acknowledge that the story does not actually begin on February 22, 2013. It begins more than 10 years ago with the development of institutional repositories; this is the existing infrastructure, disciplinary repositories, the momentous passage of the NIH public access policy, and the creation of PubMed Central. Universities and libraries have made the case for public access to research output for a long time, and librarians, in particular, have supported such campus-based activities as manuscript deposit, copyright transfer agreements, and the creation of data management plans. Universities have committed to infrastructure and Internet to long-term preservation of digitized content in the HathiTrust, so this is how we got here. Why is higher education raising its hand to be a partner and facilitating public access to research output? Because we are mission driven to do so and have an interest in wanting to maintain some control over the intellectual output that it produces independent of how the OSTP mandate or any other mandate is realized or operationalized because higher education has invested deeply in its faculty, it labs, its infrastructure, including repositories, and there is growing evidence that openness enhances research and discovery.

Consistent with the OSTP directive, SHARE believes that research publications, data, and their associated metadata should be publicly accessible for reuse including text and data mining. And having fought for this, it is in our interest—and at least somewhat under our control—to optimize the systems we have on campus for tracking research funding, collecting research funding analytics for tenure and promotion, building systems for collaborative use of data, preserving the output, and complying with funder mandates because this is, in fact, the relationship between institution, researcher, and funder.

SHARE's initial development draft focused on kind of pushing out metadata standards and federating existing repositories and based on feedback from the community, which I said we received; we have refined that focus, I think, to define and look at SHARE's essential contribution to optimize things that workflow architecture from research creation to the deposit of its products. Institutions can build and participate in SHARE and, at the same time, make a variety of local decisions according to their plans and needs, but this architecture will bring repositories closer together which is, I think, why we all built institutional repositories and which will move us all forward.

SHARE's vision, while mapped to the criteria of the OSTP memorandum, does not confine itself to this one, albeit hugely important, use case for public access. There are campus-based policies; there is emerging state-based legislation. Research is global. Funding is both private and public, and, finally, crucially, a solution to the public access mandate must include research data and a way to link publications to that data, as Amy suggested. Data are complicated, and the memorandum gets that data are complicated and allows some flexibility in the longer vision towards realizing those details. SHARE aims to reduce the administrative burden on that principal investigator, that researcher who works within our institutions and with whom we share the goal of wide exposure of their efforts. SHARE is platform agnostic looking at this architecture and nonprescriptive as to local repository development. We understand that people are in different places and will make different decisions.

So, quickly, who benefits when the Academy takes control of the research that is produced? Researchers, we argue, through better analytics,

more exposure, a single point of deposit handled at the institutional level across what could potentially be a complex array of funder requirements, funders who share our mission to demonstrate the impact of their investments and who would benefit from such metrics, universities who are the recipients and managers of the funding covered by the mandate, and the public itself through reliance on standards and protocols. Information will be discoverable by third-party services and search engines.

So while this workflow architecture needs to be built, it does have existing parts, and SHARE will use and build on those practices and protocols and standards where they exist and build collaborative solutions were they do not, and this is the path we are on—with one minute left—the path we are on is a roadmap of the working groups that I have talked about toward prototype, pilot, and implementation. We are at the beginning of this process. Follow our progress, contact us, and join us. Thank you.

John Wilbanks: I am going to say things that are completely different from everyone else that has come before me. I am here representing Sage Bionetworks. We are a nonprofit biomedical research organization. We spun out of Merck in 2009 with a group of best-in-class researchers who tried to connect biological information to health outcomes, and these are the kinds of questions that we try to answer. (From slide: "How accurately can we predict if a female breast cancer survivor will develop a second tumor?") These are the sorts of questions that used to be really hard to answer that are becoming easier to answer, and what I am going to try to do is to connect this to publishing and to access and try to show you why open access is so important for us to get what we want done.

I do a lot of other things, I wear a lot of other hats, but my main job is as the chief policy officer for this organization. This is a question that we asked about a year and a half ago, which is how accurately can we predict, purely from data, if a woman is likely to relapse after successful cancer treatment? And despite having some of the world's best scientists in our group, part of the precept of our organization is that we can answer this question more accurately if we can engage a larger audience. So we ran a challenge collected directly from patients in Sweden and the UK using existing public data as a training set with cycles donated by Google using our collaboration platform that I will show you and connecting to the publication system by saying that the winner of our challenge, the most accurate model, gets grandfathered into *Science Translational Medicine* in lieu of peer review. The challenge would count as the peer review. We had over 200 teams enter from over 40 countries, and we required code sharing, which meant that the leaderboard was changing on a regular basis, and if somebody came up with a model, anyone else could grab their code and put it into their model. As a result of this, the accuracy of the statistical model jumped three orders of magnitude in 9 days. In all of this, the only incentive—there is no cash incentive—the only incentive was a publication in a high-impact factor famous journal. The winning result was impressive enough that we did not just get an article, we got the cover. This is from April. The winning model is 76% accurate, which is unbelievably improved from the existing, and the winner is not a biologist. The winner was actually the team that created the MPEG codecs at Columbia whose ideas and theories had not been particularly welcomed by their biological politics, and they actually had three publications in 2013: one in *PLOS Computational Biology*, the cover of *Science Translational Medicine*, and a preprint in *Archive*. The preprint is most interesting to me because it shows what can happen when the article is thought of, not as the endpoint, but as a Polaroid of what is actually happening in research. So the publication is an essential carrot, which is why it is important to us, but it is actually the least significant piece of the scientific method that is going on when you start to expose this.

Let us look at what we can do with the archive version of the paper. Let us look at the scatter plots. What you really want to do was say, for that given plot, I want to know the software code that generated it and the data that was fed into that software code, and, indeed, that is what we provide. You click on this and all of this is open source software, all of this is available for free; because we are a nonprofit, we sort of would like everyone to steal our stuff. And what is interesting is, you say, when you click on that scatter plot you actually get a little directed graph that says there is

the code that creates the scatter plot. There is the "R" code, and there is the data so we can zoom in on the "R" code. We actually pull up the software code itself because, again, you had to make visible all of your software. From that, we would actually like to look at the data on which that "R" code ran, and, from that, we see that it was actually multiple datasets that were normalized. We see the names of the people who did the normalization and the names of the people who curated the data.

So the actual publication in this is just a point in time. What we really want to be able to do is follow our nose all the way back through the research process and see who did what at what point. The success of this led a large cancer project, The Cancer Genome Atlas, to say, you know, we would like to have that sort of internal provenance tracking to use inside our product. This is a classic big science project out of the NIH, and they want to develop the unique genotypes of multiple common types of cancer. This is a project that has plenty of money to run their own IT, and they decided to put their code and their data into our platform because they wanted these sorts of collaborative tracking features. If you want to work on data across space and time with people that are not in your lab, that are not on your lab circadian rhythm, you need these kinds of features; and what was really interesting is that, once we put the data into the system, here is what happened within 9 months: 68 core projects developed on top of that data within 9 months, almost 250 researchers, 28 different institutions, more than 1,000 datasets creating more than 1,700 results and 18 papers in press. Alright, 9 months. There are 36 more that have been accepted since we have made these slides.

This is an incredibly powerful method to do science, and the connection of all of this to open access is that if we wanted to scale this, and everyone that worked on our platform had to ask for permission in order to upload their paper into the system and connect it to the Providence graphs of the work, we would be dead. If we think about the PDF as the proper form of the article, we cannot do this because the article is like the crust on the surface of the earth, right? It feels incredibly thick and important when that is where we are working, but if you start to expose the scientific method that leads to that result, you see how thin the crust really is and how important it is to render all of the parts of an article clickable. Either the people doing the work can do that annotation, or we can expect that to be an incredibly expensive service provided by publishers. Alright? This is not cheap to do. I am not suggesting it is, but the people who have the motivation to do it are the researchers who get those little microattributions and microcredits; and if at every point they have to ask for permission to add links to the article or permission to make a transformation of the articles format or permission to copy it into an environment like ours, this system will not scale.

I am not up here because I want to talk about open access as an advocate—although I have done that in my past—I am up here because we want this sort of thing to be at the core of data-driven scientific research, right? The ability to copy the article into another place in the right format and add information to it and integrate it into a data centric workflow RLOA is the only answer. We have been able to do this using the articles that we have got available because one of them was in PLOS and one of them was in Archive, but if we wanted to take this to scale in cancer research, in climate research, in social sciences, in political research, any of the places where data and statistics are beginning to drive decision making to justify the choices that we make, it is only going to work if the authors and the researchers have the rights to do this themselves. Because it is already going to be expensive enough to simply store copies of the articles. It is already going to be expensive enough to keep the best available version. If we have to put the pressure on the publishers and the societies to do all of this dense integration into the workflow, it is not going to happen. There is not enough money to do simply the archiving. There is not going to be enough money to do this kind of dense integration. We have to allow the researchers to do this themselves, and we have to allow this to be integrated into the pedagogy. So I will stop there because I know we have probably quite a few questions for the group and, again, thank you, Alicia, for the chance to talk.