National Science Foundation
Award #1641014

Report on the First IEEE Workshop on the Future of Research Curation and Research Reproducibility

5-6 November 2016
WASHINGTON, DC

This page intentionally left blank

Report on the First IEEE Workshop on The Future of Research Curation and Research Reproducibility

Marriott Marquis, Washington, DC, USA

5-6 November 2016

National Science Foundation Award #1641014

Steering Committee

Chair: John Baillieul, Boston University

Larry Hall, University of South Florida
Sheila Hemami, Draper Labs
Michael Forster, IEEE
Fran Zappulla, IEEE
José M.F. Moura, Carnegie Mellon
Gianluca Setti, University of Ferrara
Gerry Grenier, IEEE
John Keaton, IEEE

Douglas McCormick and Kenneth Moore, rapporteurs

Contents

Attendees

SIMON ADAR, Code Ocean

JACK AMMERMAN, Boston University

JOHN BAILLIEUL, Boston University

BRUCE CHILDERS, University of Pittsburgh

SAMIR EL-GAZALY, University of Arkansas

MICHAEL FORSTER, IEEE

AMY FRIEDLANDER, National Science Foundation

GERRY GRENIER, IEEE

SUNIL GUPTA, IEEE

LARRY HALL, University of South Florida

KAREN HAWKINS, IEEE

SHEILA HEMAMI, Draper Labs

LINA KARAM, Arizona State University

JAMES KELLER, University of Missouri

JOHN KEATON, IEEE

LISA KEMPLER, MathWorks

PATRICIA KNEZEK, National Science Foundation

JELENA KOVAČEVIĆ, Carnegie Mellon University

MIRIAM LESSER, Northeastern University

CLIFFORD LYNCH, Coalition for Networked Information

DOUGLAS MCCORMICK, Runestone Associates

SHEILA MORRISSEY, Portico

ELEONORA PRESANI, Elsevier

BERNARD ROUS, ACM

GIANLUCA SETTI, University of Ferrara

GAURAV SHARMA, University of Rochester

EEFKE SMIT, STM International

DAVID SMITH, IET

MATT SPITZER, Community of Science

VICTORIA STODDEN, University of Illinois, Urbana-Champaign

MIKE TAYLOR, Digital Science

TODD TOLER, John Wiley & Sons

JENNIFER TURGEON, Sandia National Labs

DAN VALEN, Figshare

SHAOWEN WANG, University of Illinois, Urbana-Champaign

ERIC WHYNE, Data Machines Corp.

FRAN ZAPPULLA, IEEE

Preface

The introduction to the report that follows identifies 1993 as the birth date of the World Wide Web. This was the beginning of the end of print-based publishing that had been the bedrock of archival communication for more than a thousand years (if we agree that printing first appeared with the publication of the Diamond Sutra in 868). Throughout most of the history of print-based publishing, small groups of people had the financial resources to print and distribute what was published. Those doing the writing in early years were also few in numbers — as were those who could read what was published. That changed in industrialized countries in the 19th and 20th centuries when the value of universal literacy began to be recognized, and by the early 20th century, the printed word had become a very large industry in Europe and North America. Rapid growth of the scholarly research sector of publishing was delayed somewhat, but when higher education adopted the "publish-or-perish" imperative in the mid-20th century, academic journals and books also became big businesses. Since 1993, the digital landscape — from personal devices to the internet — has given rise to cataclysmic change in communication in general and publishing in particular. In the domain of research and scholarship, there are now heightened expectations of what should constitute the research record.

We are now at a point in the history of intellectual discourse when it is necessary to shore up and reaffirm the foundations of scientific inquiry. There is a new sense of urgency in maintaining rigorous procedures for verificiation and validation of scientific knowledge. This report is a step in an ongoing exploration of the enhanced research reproducibility that derives from new forms of research curation. The thoughtful dialogue that occurred at the workshop is reflected in the report, and I am grateful to the participants and the rapporteurs for what's been achieved. I am equally hopeful that the conversation among those who create and curate will continue, and I look forward to the next Workshop on Research Curation and Research Reproducibility.

— John Baillieul
Arlington, MA
March 2017

Executive Summary

This report describes perspectives from the Workshop on the Future of Research Curation and Research Reproducibility that was collaboratively sponsored by the U.S. National Science Foundation (NSF) and IEEE (Institute of Electrical and Electronics Engineers) in November 2016. The workshop brought together stakeholders including researchers, funders, and notably, leading science, technology, engineering, and mathematics (STEM) publishers. The overarching objective was a deep dive into new kinds of research products and how the costs of creation and curation of these products can be sustainably borne by the agencies, publishers, and researcher communities that were represented by workshop participants.

The purpose of this document is to describe the ideas that participants exchanged on approaches to increasing the value of all research by encouraging the archiving of reusable data sets, curating reusable software, and encouraging broader dialogue within and across disciplinary boundaries.

How should the review and publication processes change to promote reproducibility? What kinds of objects should the curatorial process expand to embrace? What infrastructure is required to preserve the necessary range of objects associated with an experiment? Who will undertake this work? And who will pay for it? These are the questions the workshop was convened to address in presentations, panels, small working groups, and general discussion.

Themes

Some general themes emerged during the workshop. These include:

- Funding research reproducibility and research curation activities is a major challenge in a time of flat budgets for government grants.
- Establishing research reproducibility and improving curation of research objects will have widespread benefits to science, reducing wasted effort and accelerating discoveries.
- Enhanced curation can have economic benefits as well, insofar as it leads to faster commercialization of research and new product innovation.
- Defining key terms, enunciating reproducibility goals, and setting standards for advanced processes are necessary steps toward success.
- Collaboration among stakeholders, sharing knowledge, and initiating pilot programs are the *sine qua non* of progress in this field.
- Storing objects with sufficient information to allow subsequent researchers to use them is essential. Code must be stored with sufficient information to recreate the runtime environment, or with virtual-environment wrappers. Data must be accompanied by information on how it was produced and how to interpret it.
- Making research artifacts storable, discoverable, and citable as primary objects, just as journal articles are now, is necessary to allow efficient reuse and provide incentives for researchers to do the additional work required to submit them.
- Reforming how research productivity is measured, especially in the tenure and promotion process, is necessary to encourage submissions for reproducibility review, and to reduce incentives for researchers to hoard data and code.

- Increasing the pool of reviewers for large-scale artifact review will be difficult, perhaps impossible, in an already over-stressed peer-review system. Digital tools and partial review automation may be necessary.
- The publishing model should be reformed to move away from its current focus on the printed page and its electronic images, and toward a distributed digital model emphasizing a much broader variety of research objects, linked and indexed via greatly improved metadata.
- Addressing the tension between open access and proprietary interests is necessary to ensure access to third-party objects that contribute to the research process, but are not part of the main work. Ignoring this tension may limit full participation by industrial researchers.
- Identifying and participating in relevant existing activities that provide productive guidance for further work.
- Improving software quality in research is a prerequisite of reuse.
- Addressing the challenges of version control requires some planning and standardization; and
- It is wise to keep in mind some caveats, including: Some work cannot be reproduced, and other work doesn't need to be. Fraud is rare. There is a continuum between reproducibility and reusing results. Requiring excessive levels of reproducibility review could impede science rather than help it.

Actions

The workshop plenary sessions and working groups identified actions that could accelerate progress toward wider reproducibility review and enhanced curation of varied research objects. These include:

- Research communities, publishers, funders, and others should continue to seek a viable long-term business model for reproducible research.
- Research communities, publishers, funders, and others should collaborate to establish definitions as the foundation for continuing discourse. The work that ACM (Association for Computing Machinery) has already done offers a useful starting point.
- Publishers, with research communities, funders, and other stakeholders, must begin meeting now to develop standards, particularly standards for maintaining software.
- Stakeholders need to begin selected pilot projects that build from the bottom up rather than the top down. Groups in each research community should advocate for forms of content that are more easily reproducible, and establish standards for content. Communities within IEEE and ACM that deal with particularly diverse content might be good places to start.
- The community should work to increase participation by industry in conversations about reproducibility.
- Workgroups were asked explicitly, "What actions can participants in this workshop take?" Responses included:
    - Capture as much of the reproducibility process as possible;
    - Document how different communities define reproducibility;

  - Identify several technically diverse research communities for pilot projects;
  - Identify what reproducibility programs are already underway;
  - Make reproducible results more visible (and desirable) in our own communities;
  - Establish high-profile initiatives with goals to increase awareness;
  - Encourage and reward reproducible research, without trying to coerce it;
  - Establish groups of volunteers from among workshop participants and their colleagues (either ad hoc or assisted by the workshop organizers) to work on these issues through the coming year, building to a second IEEE workshop; and
  - Create a central clearinghouse for sharing reproducibility information between communities.
- Research communities and funders should begin a collaboration to devise incentives for making research reproducible and research artifacts more accessible. Misaligned incentive structures (such as rewarding sheer numbers of papers resulting from a funded research project) can be counterproductive. Authors going the extra mile for reproducibility review must not be discouraged or shamed by criticism during review.
- The leaders of all stakeholder communities, including leading research organizations and universities, must work together to create and evaluate options for changing the way we measure the impact of research.
- Publishers and librarians, with research communities, funders, and other stakeholders, must begin to coordinate and develop broader metadata to accommodate new objects, including data, code and workflows.
- In the near term, before infrastructure is developed, publishers must be ready to host and serve artifacts from their own sites, if necessary. It is advisable for publishers to collect relevant artifacts associated with published papers — even if they are only kept in a dark archive. Third-party repositories should also be recommended by publishers for housing research artifacts, even as repository standards evolve.
- The transition to treating software and data as primary research objects should be made first in small, enthusiastic communities.

This report provides a summary record of the workshop proceedings and concludes with appendices that direct readers to additional resources for closer examination of aspects of scholarship curation and reproducible research.

Introduction

It has been less than a quarter of a century since the creation of the World Wide Web was announced at CERN, the European Organization for Nuclear Research, in April of 1993. Since then, concepts of research curation and research dissemination have undergone greater change than in the previous 600 years. Change has been so rapid and dramatic that agencies concerned with organizing the newest additions to the large corpus of human knowledge have struggled to keep requirements and standards relevant to the current practices of the research community.

This report describes perspectives that emerged at a workshop on the Future of Research Curation and Research Reproducibility that was collaboratively sponsored by the U.S. National Science Foundation and IEEE in November 2016. The purpose was to conduct a deep dive into new kinds of research products and how the costs of creation and curation of these products can be sustainably borne by the agencies, publishers, and researcher communities represented by workshop participants. This document describes the ideas that workshop participants exchanged on approaches to increasing the value of all research by encouraging the archiving of reusable data sets, curating reusable software, and promoting broader dialogue within and across disciplinary boundaries.

To be of maximum value, the results of scientific research must be saved (stored and preserved), shared (accessible, discoverable, and citable), and trusted (comprehensible, reviewed, and reproducible).

Modern research generally, and research in engineering and computer science in particular, depends heavily on computation. Research results include not only the output data and analyses as presented in traditional scientific papers, but also new kinds of results that present new challenges to the save-share-trust process. These results (referred to here as objects or artifacts) include large-scale input and output data, software, detailed experimental methods, and enough information on the computing environment to allow other researchers to reproduce the experiment.

There is growing interest in ways that the value of research will be enhanced by new and rapidly evolving forms of curation. Curated research artifacts will have value to the extent that

- They are easily discoverable;
- They allow experiments to be repeated to determine whether published results can be reproduced;
- They enhance the usability of data and software for novel purposes.

To assure that new research will have this added value, funders, publishers, and researchers must develop new approaches to the review-and-publication process. Questions addressed in the workshop included the following: How should the review and publication process change to promote reproducibility? What kinds of objects should the curatorial process expand to embrace? What infrastructure is required to preserve the necessary range of objects associated with an experiment? Who will undertake this work? And who will pay for it?

Workshop Overview

On 5-6 November 2016, the IEEE, supported by National Science Foundation Award #1641014, convened a workshop of 37 expert stakeholders to consider how these questions might be addressed to improve the publication of engineering and computational science. Workshop participants included representatives of funding organizations, the research community, for-profit and not-for-profit publishers, academic research libraries, and computer services evolving to support preservation and reproduction of nontraditional research objects such as large data sets and software, and others.

The workshop included two introductory presentations and three panels. Each panel addressed one of three main questions (see appendix for expanded versions of the three main questions):

- What are the essential products of scholarly engineering research, and how will these be likely to change in the future?
- How can interrelated, constantly updated research products — including experimental protocols, archived data, software, and other nontraditional artifacts, as well as traditional scholarly papers — be reviewed, stored, maintained, searched, accessed, and cited.
- What are economically sustainable approaches to providing public access to engineering research products?

Each panel was followed by breakout sessions, in which working groups of 9-10 participants addressed assigned aspects of the panel's main topic. The first day of the workshop concluded with an opportunity for each of the participants to summarize what he or she saw as important messages to take away from the meeting.

This report presents summaries of panelists' presentations and breakout session comments. All summaries are paraphrased, unless direct quotation is indicated. The comments express the views of individual participants and do not necessarily reflect the opinions of the workshop as a whole or its organizers.

Themes and Actions

Some general themes emerged during the workshop. The most-often raised were problems in funding research reproducibility and research curation, the need for definitions of key terms and standards for advanced processes, the necessity of collaboration, the importance of making research objects discoverable and citable, the need to reform how research productivity is measured, and the difficulty of supporting large-scale artifact review. Only slightly less prominent were calls for (a) shifting publishing models from their current focus on the published page to a distributed digital model, (b) recognition of the ways that reproducibility and curation would benefit science as a whole, and (c) addressing the tension between open access and proprietary interests. There was also substantial discussion of the importance of developing metadata models, existing activities that can provide guidance for further work, the role of software quality engineering, and the challenges of version control. Finally, it is worth noting a series of caveats warning against thoughtlessly embracing grand but unsustainable plans for achieving reproducibility in research results. **(Potential actions discussed in plenary sessions and workgroups are boldfaced.)**

Funding

Research funding is currently flat. Money allocated to validating reproducibility reduces the pool for traditional research products. National Science Foundation Deputy Division Director Amy Friedlander stressed this constraint in both opening and closing remarks. Most segments of the research enterprise face the challenges of a zero-sum game. Libraries must allocate funds among content, services, and infrastructure, enhancing any one area by reducing support to the others. More broadly, researchers, libraries, and publishers compete for the same fixed pool of money.

The greatest challenge to developing reproducible and curated research results lies in devising a sustainable economic model, and, likely, in deciding which functions of traditional publishing must change radically or disappear to make room for improving outputs and processes.

Workshop participants considered whether any of the stakeholders in research publication clearly could or should fund reproducibility review and curation. Should funders cover the costs, through grant set-asides, or through reallocation of institutional overheads? Again, though, increasing funds in one area means reducing them someplace else. Should publishers pay? Perhaps, but then the costs would clearly be passed on to researchers and libraries, though radical (mostly technological) reforms in publishing, reviewing, and curation might yield efficiency savings that could cover the costs of enhanced artifact review and curation. Should industry pay, since it reaps benefits from research? Perhaps, but it was unclear how to motivate, or even compel, corporate contributions commensurate with commercial benefit. Should an industry consortium or a non-profit foundation underwrite the effort? But if they did, would awards be long-lived enough to be sustainable? Might researchers pay, via some sort of crowdfunding mechanism that would tie reproducibility awards to the level of peer support in the research community?

For the issue of funding, among the thorniest questions of the two-day workshop, there were no easy answers. Workshop participants agreed that the topic needs continued exploration – possibly in a follow-on workshop.

Action: Research communities, publishers, funders, and others should continue to seek a viable long-term business model for reproducible research.

Definitions

The vocabulary of reproducibility currently lacks fixed technical definitions. Reproducibility, repeatability, replicability, verifiability, reusability, validation and other key concepts mean different things in different disciplines, and are often used interchangeably. The issues go beyond reproducibility alone: to be truly useful, research artifacts of all kinds must be shared, preserved, accessible, discoverable, citable, comprehensible, reviewed, reproducible, and reusable.

The Association for Computing Machinery (ACM) has already debuted an effort to define terms, based on formal International Bureau of Weights and Measures (BIPM) usage, and to rate the degree of reproducibility of the papers it publishes.

Action: Research communities, publishers, funders, and others should collaborate to establish definitions as the foundation for continuing discourse. The work ACM has already done offers a useful starting point.

Standards

Definitions and standards go hand in hand. Standards are needed for assessing reproducibility, submitting papers and artifacts, and for preserving, tagging, curating, and citing artifacts. Standards will evolve over time, but it is important to begin now to draw in all stakeholders and open broad standards development.

Action: Publishers, with research communities, funders, and other stakeholders must begin meeting now begin to develop standards, particularly standards for maintaining software. Intellectual property (IP) considerations, privacy concerns, and sheer variety make it much more difficult to develop standards for collecting, maintaining, reviewing, and serving data objects.

Collaboration, Sharing Knowledge, and Pilot Programs

A consensus is beginning to form that future reports of experimental research will include access to the artifacts necessary to reproduce it. Making this the norm will require collective action over a very broad and diverse front; no single group can do the job. In the past, each publisher and each library has worked alone, trying to address even common problems in isolation. This is beginning to change, with the emergence of regional, national, and industry-wide consortia and collaborations.

Enthusiasm for ensuring reproducibility and norms for peer review vary among technical disciplines and subdisciplines. Lessons can be learned and allies recruited from established data repositories and other collaborative initiatives, such as the Inter-university Consortium for Political and Social Research (ICPSR) for social science data and GenBank for genomic data; active collaborative software efforts, such as GitHub and the open software community generally; more formal organizations, such as NISO (National Information Standards Organization); RMap; CrossRef and CrossMark; ORCID; CREDIT; Hypothes.is; the Center for Open Data Enterprise; RDA (Research Data Alliance); CODATA (the Committee on Data for Science and Technology of the International Council for Science); the Illinois Data Bank, a regional public access repository at the University of Illinois at Urbana-Champaign; and emerging artifact and digital-first journals, from Elsevier, John Wiley, and the Public Library of Science (PLOS), for example.

In ACM's review-of-reviewing effort, the association is drafting (a) best-practice guidelines for data, software, and reproducibility review, (b) an XML (Extensible Markup Language) schema for artifact metadata, and (c) a legal framework. All may be released during 2017.

Action: Start somewhere, and start soon, to begin selected pilot projects that build from the bottom up rather than the top down. Establish groups in each research community to advocate forms of content that are more easily reproducible, and establish standards for content. Communities, especially within IEEE and ACM, that deal with particularly diverse content may be good places to start.

Action: Increase participation by industry in conversations about reproducibility.

Action: Workgroups asked, "What actions can participants in this workshop take?" Responses included:

- **Capture as much of the reproducibility process as possible;**

- **Document how different communities define reproducibility;**
- **Identify several technically diverse research communities for pilot projects;**
- **Identify what reproducibility programs are already underway;**
- **Make reproducible results more visible (and desirable) in our own communities;**
- **Establish high-profile initiatives with goals to increase awareness;**
- **Encourage and reward reproducibility, without trying to coerce it;**
- **Establish groups of volunteers from among workshop participants and their colleagues (either ad hoc or assisted by the workshop organizers) to work on these issues through the coming year, building to a second IEEE workshop; and**
- **Create a central clearinghouse for sharing reproducibility information between communities.**

Citation, Discoverability, Incentives, and Credit

If artifacts are to be used, researchers must be able to find them and guide others to them. If artifacts are to be submitted, researchers must have incentives to do the additional work needed to prepare and review them. (By the same token, publishers need to streamline procedures and provide tools that lower labor barriers to submission as much as possible.) Building a culture of reproducible research requires that software and data artifacts can be stored, cited, and searched as primary research objects — to "live their own lives," as one workgroup put it — just as traditional papers are now.

Researchers are motivated by funding and scientific credit — fortune and glory, if you will — in addition to other factors. In June 2015, ACM began crediting artifact reviewers and opening its journals to articles detailing the results of the artifact review. (Note, however, that proprietary interests may legitimately prevent researchers from making software and data available for review — just as industrial research is frequently never published in the open literature.)

Workgroups concurred that an essential element in motivating researchers to submit or review data and code is ensuring an incentive structure that includes kudos or credit for their work, though it is unlikely that one can build an academic career at a top-tier research institution by reproducing others' research: the resulting papers will lack novelty almost by definition.

The indices used to measure scientific productivity desperately need reform, particularly in the promotion and tenure process. Usage metrics for open data and software publications are beginning to appear in tenure and promotion packages. And at least one NSF "committee of visitors" was surprised to find that some junior faculty are submitting review-panel critiques of rejected funding proposals as part of their tenure packages, because so few applications are accepted. As the products of research outgrow the traditional scientific article, are there better ways of demonstrating research productivity to tenure and promotion committees, the university at large, and funders and policy-makers?

Action: Involve research communities and funders in devising incentives to perform and submit reproducible research. Work is needed to create and evaluate options for changing the way we measure the impact of research.

Reproducibility Review Is Not Scalable

Reviewing is highly skilled, time-intensive, unpaid labor. Artifact reviewing is even more labor-intensive, and likely calls for different sets of skills. Adding a full artifact-review mechanism on top of traditional article peer review may well call for more resources than exist. And making the artifact review process too onerous for author or reviewer will just impede scientific progress. Postdoctoral fellows and graduate students may be willing and able to take on part of the work, but many of them, though scientifically proficient, lack the necessary written communication skills. Standards of review may have to evolve to include some degree of automation and the introduction of non-blinded reviewing that can result in independent publications.

Workshop participants suggested alternative ways of conducting and funding artifact review: establishing intramural repositories in which validation can begin before publication (and, indeed, before submission); prepublication reviews leading to simultaneous publication of artifact review and primary article; post-publication review (possibly crowdsourced, though the 15-year record of open annotation is not encouraging). Some disciplines (such as crystallography and cytometry) have already developed workflows for pre-publication data validation, which might offer models. If artifacts are archived in a form that supports peer review, the papers based on them may tend to be easier to understand.

Reproducible Science Is Good Science

Reproducible science is good science. Increasing the trustworthiness of published research reduces effort wasted in building on faulty science. Recovering and running software from a published paper can be the work of many months. The greater availability of code and data, a byproduct of reproducibility analysis, makes follow-on research easier and more efficient. Giving credit for software and data might, in the view of some, curtail hoarding of data and code to spin out as many (possibly low-quality) publications as possible. It is thus conceivable that a move to reproducibility could promote publication of fewer, higher-quality publications.

Amy Friedlander stressed that research projects are getting bigger, increasingly multidisciplinary, and more dependent on computation. Reproducible and accessible science is good science, she said, giving a firmer foundation for future work.

Digital First

As one workshop participant put it, “We’re doing a lot of analog thinking in a digital world.” Publishing and data preservation paradigms require fundamental reform. Even today, publishers still think in terms of the printed page, taking electronic files and “stripping out their value” to typeset them and produce PDF (Portable Document Format) facsimiles of traditional journals. Publishers must rebuild workflows from the ground up, toward the kinds of objects — JSON LD (JavaScript Object Notation for Linked Data) and Schema.org — that enhance scientific meaning and are intelligible to Google and the open web.

Libraries, too, are responding with radical reviews of how they fulfill their functions, becoming the “library as platform.” Some are adopting just-in-time acquisition policies, purchasing journals and books when they are requested by researchers. Others are radically reviewing staffing and resource allocation, even cutting into time-honored functions like the circulation desk. Many are

moving toward regional consortia and integrated web services that allow access to holdings far beyond individual institutions.

As research becomes increasingly computational, multi-disciplinary, and (often) global in scope, there may come a time when artifacts like code and data are valued more highly than printed descriptions of their operation and production.

Access and Proprietary Issues

Reproducibility analysis requires that reviewers have access to the researcher's original code and data. Some visions of reproducible science include even more widespread access. There is a tension between open access, as increasingly mandated by government funding agencies, the proprietary copyright control underlying traditional scientific publishing, and the intellectual property rights of researchers and suppliers of third-party software.

Intellectual property policies have the power to shape how the reproducibility and curation discussion proceeds, what programs are implemented in the community, and what infrastructure is built. Once these are set, it will be very difficult to change course. Even today, researchers who hold all rights to their own code may be unable to submit it for thorough review because the work relies on third-party software that they cannot redistribute. One of the key steps in establishing the Defense Advanced Research Projects Agency (DARPA) Open Catalog software archive was to create a "sensible" license structure.

At the same time, U.S. research universities pay more attention to intellectual property than they did 20 years ago and much more than they did 50 years ago. For these and other reasons, tensions may arise unless there are open and frank discussions within the research community about how open access to software, especially the products of publicly funded research, really should be — to strike a balance between respect for proprietary rights and reproducibility. It is important to understand when objects can legitimately be proprietary and when they should be open. It may be necessary, though difficult, to separate questions of replicability and reproducibility from questions of open data and open-source software. The experiences of the ACM, where industry and academia meet and overlap, teach important lessons about accommodating these relationships in our thinking about replicability and reproducibility.

The obstacles to sharing data may be even greater than they are for software. Researchers often use others' data, but there are privacy concerns in addition to intellectual property restrictions that may prohibit them from sharing it further. Moreover, evaluating data requires a different set of skills and different incentives.

Metadata

Metadata schemas are essential to archiving, accessing, and citing research artifacts. Even before new publishers and repositories can develop new infrastructure for housing and serving artifacts, well-designed metadata databases can make scattered artifacts searchable and accessible. (ACM and others are currently developing metadata schemas for their disciplines.)

Action: Publishers and librarians, with research communities, funders, and other stakeholders, must begin now to coordinate and develop broader metadata to accommodate new objects, including code and data.

Tools

It will take time to develop the tools needed to routinely submit, review, store, and serve artifacts. Burdens on researchers may well be unacceptably heavy unless they have access to automated tools for preparing and submitting software, for example. Reproducibility reviewers will face unacceptable challenges in recreating runtime environments and debugging their own installations of software for review, unless they have access to "wrappers" — encapsulated runtime environments. This does, however, present legal and technical challenges. Ultimately, well-indexed links to a variety of artifacts should allow readers to launch simulations and programs seamlessly from within an online article, run code with their own data, and modify code within the online environment.

Action: In the near term, before infrastructure is developed, publishers must be ready to host and serve artifacts from their own sites, if necessary. It is advisable for publishers to begin now to collect software associated with published papers, even if in a dark archive.

Software Quality Engineering

Risk-based quality control (such as practiced at Sandia National Laboratories) is essential in projects that involve possibilities of great economic harm, or danger to human health and life. Standards for quality engineering are necessary for such projects, and desirable for a wide range of others. Peer review and researcher integrity are cornerstones of the process, however much project scales or budgets might vary. When it comes to the details of the review, one size emphatically does *not* fit all, and elements must be appropriate to the project's risks, scale, and budget.

Participants discussed differences in software development and maintenance practices between software engineers to the point that, in one computer scientist's experience, well-intentioned efforts by software engineers to "clean up" code — to make it more robust for use by other researchers — actually had the opposite effect.

Versioning

Versioning is a perennial challenge. Sandia National Laboratories archives software whenever it reaches a break- or release-point. Utilities like GitHub also offer secure archiving of interim versions. Similar version archives should be established for third-party software used in projects, but protocols there are yet to be developed. Data sets, too, should be versioned and archived, along with information on the rationale by which the raw data is filtered and processed to become input.

Caveats

Clifford Lynch, executive director of the Coalition for Networked Information (CNI), outlined important financial, regulatory, and practical constraints on the push for better curation and increased reproducibility of research results. In particular, some papers cannot be reproduced (and others don't need to be); fraud is rare; and there is a continuity between reproducibility and reuse of results. It is easy to *say* that artifacts must be deposited as part of publications. It is quite another thing to ensure that those deposits are made. Hypertrophied, "fetishized" peer review had become a barrier to effective scholarly communication until the emergence of public preprint

archives, such as arXiv. Effort misapplied to excessive review could be characterized as a profligate waste of human resource.

ACM has recommended that publishers should *not* require confirmation of reproducibility as a condition for accepting a paper. Lynch noted it would be counterproductive to build expensive and slow replicability processes into the publication system itself. Sheila Hemami asked, "How do we allocate precious peer-review resources? And how do we make sure that material not selected for detailed review can nonetheless be published, preserved, and used as it merits?"

ACM also suggested that artifact reviews might not be blind, but rather open collaborations with the author, possibly resulting in an independent publication by the reviewer.

Raw data is generally not data actually used. It is not enough to simply cite public data or publish raw data files. Authors must also detail how the data were filtered, and cleaned up to yield the published results.

Action: The transition to treating software and data as primary research objects should be made first in small, enthusiastic communities.

Introductory Presentations

Public Access, Data, and the Research Enterprise

Amy Friedlander, Deputy Division Director of the Division of Advanced Cyberinfrastructure in the Computer & Information Science and Engineering Directorate of the National Science Foundation.

Friedlander presented the research funder's view, beginning with the research context at the National Science Foundation, and particularly at the Division of Advanced Cyberinfrastructure, which grants funding to support future science and engineering.

Overall, some 94% of the $7.3 billion allocated to NSF for FY 2019 will be distributed in awards to 1,900 institutions, generally after competitive merit review. This means that the foundation runs on "what is effectively a 6% overhead." The agency runs lean and uses the sorts of information tools it helps develop to streamline its own processes, while making sure they continue to apply to the very wide range of research funded by NSF's seven directorates and seven main offices.

The trick to all of this is to take this structure and advance science at a time when interdisciplinary research is becoming more and more technology-complex, as well as complex from a disciplinary perspective.

"The trick to all of this is to take this structure and advance science at a time when interdisciplinary research is becoming more and more technology-complex, as well as complex from a disciplinary perspective," Friedlander said. Research regularly crosses disciplinary boundaries. The question once asked about biochemistry — is it biology or is it chemistry? — pales in comparison to the disciplinary complexity of modern computational molecular biology, with its foundations in computer science, biochemistry, biology, and physics.

Projects funded by Advanced Cyberinfrastructure aim at provisioning the entire research enterprise. When the division commissions an advanced supercomputer or an advanced network, the goal is to show what research will look like in a decade or two, rather than to simply build capacity.

Friedlander presented summaries of four projects that illustrate the breadth of technologies used in the division's research and the equally wide range of uses that its results support:

- A project to develop an interactive, empirical map of America's food, energy, and water systems to model impacts of economic production, consumption, and agricultural trade; political, economic, and regulatory stresses and shocks; water systems; environmental flows; carbon dioxide emissions; and land use.
- An "Array of Things" project started with the deceptively simple objective of distributing 100 to 500 varied sensors (for light, air pressure, gas levels, traffic, sound, temperature, etc.) in an urban environment, and then collect and organize the data into a framework that would support research in public health, climate, weather forecasting, and social dynamics.

- A computational infrastructure for brain research, a cloud-based experiment-management system to support open data sets and analysis from neuroscientists worldwide.
- Development of a real-time network monitoring instrument to understand how network traffic changes over time, detect unauthorized attacks, spot network inefficiencies, and understand the behavior of both human and automated traffic.

Data Management and Access Plans

NSF prides itself on having had a data management/data sharing policy that goes back to the 1990s. The original draft seems ambitious today. It covered publications, software inventions, data in all formats, and intellectual property. It called for decentralized administration, recognizing that the highly heterogeneous scientific and engineering research NSF funds yields highly heterogeneous results. The products include everything from instrument calibration tables to curriculum development to mega-projects like the Large Synoptic Survey Telescope, the Large Hadron Collider, and Litho.

Beginning in January 2011, NSF started requiring that new grant applications include data management plans in their proposals. Two years later, in January 2013, the agency began to "allow and encourage" applicants to include citations to data sets in their application biographical sketches, and to report their data in annual and final reports.

In a far-reaching memo of 22 February 2013 (the Holdren Memo)[1], the U.S. Office of Science and Technology Policy (OSTP) required public access to all results of federally funded research, including journal publications and data.

As part of the development of its existing Data Management Plan, and also in response to the Holdren Memo, NSF issued an 18 March 2015 Public Access Plan.[2] The plan calls for a federated architecture that builds on the agency's existing data management plan, accommodates the wide variety of research, research cultures, and outputs of NSF-funded research, leverages distributed infrastructure, resources, and services, and *requires the agency to consult with stakeholder communities* to develop guidance and to conduct pilots in key areas, notably in the definition, creation, distribution, and maintenance of identifiers. (Asked later how "community" should be defined in this context, Friedlander replied that, within broad boundaries, communities are self-identified by active members coming forward to participate in the process.)

Seven months later, during October through December of 2015, the NSF Public Access Repository (NSF-PAR) went live. Now, principal investigators could deposit and report on their research through research.gov. NSF program directors could review progress and evaluate reports through the agency's in-house *eJacket* system. And the general public — including the research communities — could locate and access research results via https://www.fastlane.nsf.gov/ and par.nsf.gov. (The first submission to NSF-PAR arrived, unsolicited and independent of any pilot program, less than a week after the system debuted.)

The evolving PAR architecture is community-driven and extensible, with identifiers, metadata, and standards for storing objects in either distributed or centralized architecture. The open questions are: Will this architecture work for data and software? Under what circumstances? And

in one system or many? To answer these and other questions, NSF made 26 grant awards in FY 2014-2016, to study the issues and develop revised guidance and tools for public access.[3]

"[The 2015 Public Access Plan] obligated the agency to engage in a sustained conversation with the research communities to develop new guidance or revised guidance for the directorates in the implementation of the data management plan," said Friedlander. "That is equally important as … standing up a repository for journal publications and juried conference papers."[a]

Beyond the practical and technical issues of implementing data management plans are the more important questions of how these new structures will affect the science. "[K]eep in front of you as you work through the issues that will be before you, that this is about science," Friedlander told the workshop. "This is not about regulations. The regulations are there in order to ensure responsible management and conduct of the work. Ultimately, we do this in support of science and engineering. … [T]o reiterate: we believe that data management is good science. … The question is, can we codify it in such a way that it becomes shareable with others …" The solutions must stretch to include both large and instrumental projects, but also address the very labor-intensive problems posed by smaller legacy collections.

Overview of the Reproducibility Landscape

Clifford Lynch, Executive Director of the Coalition for Networked Information (https://www.cni.org/).

Lynch began his presentation by stressing three principal points:

- *Some papers cannot be reproduced. Others need not be.* An observation of a one-off event, for example, "is what it is, and it's over with. It's important to put that caveat very squarely because it's too easy to get obsessed about reproducibility" and impose inappropriate blanket requirements.
- *Fraud is rare*. "Very few researchers sit around trying to think about, 'What can we fabricate … and publish?'" Much more common are honest errors, or over-enthusiastic or over-helpful interpretations and analyses of data:

> While it's important to deal with outright fraud, that's not the main purpose of the system. The main purpose of the system is to do better scholarship. Everybody has a stake in that. The scholars themselves, the funders, the public policy folks, the publishers, everybody has reasons to want to do sound scholarship. That's particularly important, by the way, in areas that have high stakes, a notable example being biomedical research where, if you think about it, a certain kind of effort to ensure reproducibility has actually been codified in the approval of drugs; the entire clinical trial process is a very formal way of putting things under scrutiny. I'd also

[a] Managing publications, Friedlander noted, becomes a "special case" of the larger data-management challenge. Publication management is robust and decentralized, using distributed infrastructure capabilities that include CrossRef (crossref.org), the ISSN (International Standard Serial Number) identifiers, [Chorus ("Manage your spectrometry files online," https://chorusproject.org).]

> suggest that many aspects of engineering represent high stakes work. It's important, I think, to recognize that.

- *There is a continuum between reproducibility and reuse of results*. There is no clear dividing line between checking results and using them to support further work. "I'd suggest that reuse is equally as important as reproducibility, and that over time, [reuse] perhaps has an even higher payoff in terms of accelerating the progress of scholarly work."

Several Aspects of Reproducibility

Description of processes. The "materials and methods" section of the traditional research paper has been the classic tool for ensuring that the results can be reproduced. This section is often impenetrable, often left unread, and often relegated to a smaller font, a figure caption, or online supplementary material. Now, however, methods are becoming more accessible. Technique videos, for example, may be included as part of the paper or supplementary material. There are also efforts to collect, codify, and update protocols and research methods so that they are more easily available and more easily corrected or expanded.

These elements are best collected in a form that is easily annotated. Published protocols sometimes contain errors; without a means of quick and reliable correction, these flawed methods can propagate through a field, wasting research time, money, and resources as one group after another has to discover and correct the bugs in them.

It's tempting to assume that abstract mathematical and theoretical computer science papers shouldn't have problems with reproducibility, that work based on formal proofs should be intrinsically reproducible. In fact, though, the refereeing process for these papers is (at its best) a detailed confirmation of reproducibility, as the reviewer stumbles over and corrects unclear statements or lapses in logic.

We know from hard experience that it is really, really, really hard to write correct software. Incredibly, software can be out there in wide use for decades, functioning incorrectly.

In the physical world, the tools one uses are an integral part of the results. The specific instruments, software, hardware, cell lines, etc., have a fundamental impact on the results. You can't have a meaningful result without a thorough and reproducible means of obtaining it. In many cases, the validity of an analysis is inseparable from the underlying data, and without that data there can be no meaningful confirmation. So the data must be available for the analysis to be believed.

Computational tools are integral to a great deal of current research, in engineering disciplines and in science generally. This is one of the reasons reproducibility has become a sensitive issue. Checking computation manually is almost never a practical option, so it is essential to establish that the software is reliable. Yet, as Lynch observed, "we know from hard experience that it is really, really, really hard to write correct software. Incredibly, software can be out there in wide use for decades, functioning incorrectly. People find these errors all the time." And as software becomes increasingly layered and complex, there are more and more opportunities for errors to creep in. Hardware can also malfunction or be designed with errors.

Thus, if we are to really understand the relevance of a research result, we should be able to re-check the computations, the software, to understand if and how it may go wrong. This boils down to two separate problems:

- "Can I preserve the whole software environment that produces these results, so someone else can get them again?"
- "Can we understand the software's sensitivities, so that when someone discovers an error in a crucial piece of software, you can walk it back and find out what various results in the literature relied on the correct operation of that software, and were potentially tainted by downstream errors?"

These concerns apply to all the software used in the work, not just the routines developed for the project. They apply, rather, to everything from source code to third-party tools, both open-source and proprietary commercial packages.

Intellectual property rights aside, it is technically possible to emulate an entire computational environment. As Lynch noted, "Operating system, libraries, applications can all be packaged together into what's in essence a virtual machine or a container, which can be stored with the research as connected to the research report, and then reanimated by anybody who wants to run it."

At the same time, formidable problems of copyright, standardization, and scalability remain to be solved before "containerization" can become a standard practice. Moreover, while containerization "is a perfectly reasonable way to think about a fairly high-powered desktop environment, it is a much less feasible way to think about a supercomputer with a very large number of processors."

Preserving the software in its environment is not the only approach that makes sense. "There's a very interesting set of trade-offs between redoing computation and preserving the results of computation."

What is the best archiving strategy? 'Do you keep the results? Do you just keep the code, knowing that if anybody cares enough, they can burn another carload of computer time and hopefully regenerate the results? Do you do something in between?'

In simulations, for example, the code itself is often compact, but it consumes a tremendous amount of computer time and generates tremendous volumes of results. What is the best archiving strategy? "Do you keep the results? Do you just keep the code, knowing that if anybody cares enough, they can burn another carload of computer time and hopefully regenerate the results? Do you do something in between?"

There is as yet no well-established practice, and these trade-offs should probably be reassessed periodically, as economic balances shift between storage and processing time.

It's important, too, to consider when researchers should be expected to demonstrate that their research is reproducible: before publication (as part of the review process, perhaps only to reviewers), upon publication (putting the "public" in publication), or after publication (establishing a formal or de facto embargo to allow the researcher to produce additional

publications)? And if data sets are released, even on a limited basis for review, what are the risks that they will be misappropriated, and how can these risks be managed or mitigated?

Publishers are the traditional custodians of the scholarly quality assurance process. Should they therefore also set the standards for and mediate the verification of reproducibility, requiring data deposits beyond the core publication, either in-house, as part of the publisher's online supplementary material, or contributed to a third-party repository? In genomics, for example, publishers require researchers to deposit genetic sequences in the public GenBank database, and a GenBank identifier, a "proof of deposit," is a prerequisite for publication. "The publishers have acted as enforcement mechanism early on in getting that practice standardized and fully adopted."

Finally, Lynch noted:

> There's a difference here between policy and implementation. It's very easy for publishers to casually say, "We expect a commitment on the part of our authors that if other scholars want to reproduce the results they publish here that they'll make the material available." In practice, though, studies have shown that authors often fail to comply with these requirements once the paper is published.

Thus, it seems clear that someone other than the researcher needs to be a genuine guarantor that the material needed to establish reproducibility will be available. This may be a data repository or the publisher, but it should not be the author.

Plenary Panel: New Forms of Content

What are the essential products of scholarly research, how will these be likely to change in the future, and how can the results of the research be accurately reproduced? This panel will identify new types of content and the challenges of reproducibility.

Panel Moderator: Larry Hall, Department of Computer Science and Engineering, University of South Florida.

Panelists: Jelena Kovačević, Carnegie Mellon; Simon Adar, Code Ocean; Eric Whyne, DataMachines.io; Sheila Morrissey, Portico.

Reproducible Research

Jelena Kovačević, Hamerschlag University Professor and Head of Electrical and Computer Engineering and Professor of Biomedical Engineering at Carnegie Mellon University. She is a former editor-in-chief of IEEE Transactions on Image Processing.

Kovačević dates her interest in reproducibility to "Pushing Science into Signal Processing," a 2005 article by M. Barni and F. Perez-Gonzalez,[4,b] and she presented her thoughts as the editor-in-chief of the *IEEE Transactions on Image Processing* in a special session on reproducible research at ICASSP 2007.[5] Together with P. Vandewalle and M. Vetterli, she explored the issues further in a popular article.[6]

She presented two examples of how reproducibility review can work in practice. In one case, the review strengthened a celebrated success; in the other, it failed to detect a misconduct that created a scandal.

Capping a series of lectures at Cambridge University in June 1993, mathematician Andrew Wiles presented a proof of Fermat's Last Theorem, one of mathematics' chief enduring challenges. Peer reviewers found an error in the initial proof, however, and Wiles and a collaborator spent a year looking for a way to fix it. This they did, publishing the final, confirmed proof in two papers in May 1994.[7,8]

A decade later, Woo Suk Hwang and colleagues published in *Science* an equally electric paper describing the first cloning of human embryos.[9] In this case, post-publication reviews and unsuccessful efforts to replicate the experiment prompted an inquiry. Investigation found multiple flaws and fabrications in that paper and others published by the group. Ultimately, *Science* retracted the papers.[10]

The first is a mathematical paper without data or physical methods; the review was a disciplined reproduction of the author's logic, and it worked as it was supposed to. The second involved research that depended heavily on method and technique, and produced results with the usual

[b] "Replicability is one of the main requirements, possibly the most important requirement, of a well-conducted experiment. For an experiment to be credible, the experimental setup must be described very accurately so that any researcher can reproduce it and obtain the same results. If other scientists cannot reproduce your results, chances are it may have been due to a flaw in the experiment rather than a real effect." — M. Barni; F. Perez-Gonzalez, *IEEE Signal Processing Magazine*, July 2005

ambiguity of living systems. The prepublication review apparently tested the paper's logic and the plausibility of its results, but did not seek to reproduce the experiment and its results. The error was indeed caught by the greater community, but only after false results entered the literature and an unknown number of other researchers had squandered an unknown amount of effort trying to either replicate false results, or conduct further research built upon this unsteady foundation.

Engineering and computational science occupy a middle ground. Many publications include formal propositions that can be logically reproduced and verified, as in a mathematical proof. One would think that affirming reproducibility would be as straightforward as it is in math. But much of engineering and computational science work also depends on complex interactions of cybernetic and physical components, and they produce sometimes complex data, as in molecular biology.

There may still be a tension between confirming reproducibility and promoting innovation. … A paper confirming reproducibility of another is by definition not innovative. 'And unless there is really novelty in the paper, people are not going to be excited about the work.'

IEEE has made great strides in promoting reproducible research and encouraging authors to make their work reproducible by providing not only the paper but also data, code, etc. There may still be a tension between confirming reproducibility and promoting innovation. Reviewers are asked, "Is the paper technically sound? Is the coverage sufficiently comprehensive? How would you describe its technical depth and technical novelty?" A paper confirming reproducibility of another is by definition not innovative. "And unless there is really novelty in the paper, people are not going to be excited about the work," Kovačević noted.

Data presents special issues. Many researchers work on problems using others' data. The data may be proprietary. It may contain sensitive personal or health information. A project may pull together software from many colleagues or collaborators, and the infrastructure may include proprietary software from many others. Who owns what? Who can share what with whom? What is a proper description of the data?

Kovačević cited a project that uses New York City taxi information, public data that included routes travelled and time and place of pick-up and drop-off. Combining this information with, for example, restaurant receipts from sales tax rolls, provides something like a movie of resident movement and economic activity. In publication and conversation, the authors identify their inputs as "publicly available data." As with any raw data set, however, the information must be curated to remove outliers and errors, and to extract the information that is relevant to the project.

If reproducibility reviewers attempt to use the publicly available data, the raw data, to reproduce the results, they may very well fail. So for the research to be truly reproducible, the researcher must provide either a detailed protocol for processing the data, or the processed data itself. Or both. It might also be possible to use the taxi data to identify specific cabs and their drivers, violating prohibitions on releasing personally identifiable information.

Consider a collaboration with a medical school in a study of middle ear infections in children. The project aims at creating algorithms to tell apart middle ear infection that requires using antibiotics from other conditions that do not, using expert otoscopists as a gold standard. The data must be anonymized, of course, which can be tricky and is governed by strict protocols; releasing such data is often not an option.

Kovačević introduced a concept of "educational reproducibility." This approach might include, for example, a mathematical companion that includes objects — software and data — that allow students to reproduce all of the technical figures in a volume and change parameters to see how the system responds. She and her co-authors did just that with their book, *Foundations of Signal Processing*,[11] releasing a Mathematica companion that allows the readers to reproduce the authors' figures and test the reproducibility. This simple educational project includes educational components that are built on the principles of reproducibility.

Why Is Curation Important?

Simon Adar, founder and CEO of Code Ocean, a Cornell Tech incubated startup (New York, NY).

There is an exponential increase in the amount of curated research, fueled by new content delivery methods, archiving, preprints, and nontraditional types of research outputs. It demands that we reexamine our priorities about what should be curated and why it is important. If we set aside the publish-or-perish paradigm, competition, and financial drivers, shouldn't reproducibility and reuse be at the top of the list?

A number of recent studies show that we have serious issues when it comes to reproducibility (http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970) (https://osf.io/e81xl/). Part of the problem is that what we do is very complex, but a good amount of blame lies with our current curation practices. Adar experienced this first-hand when working on a graduate project to combine on-the-ground, satellite, and airborne data to detect ground pollution around mining operations. Adar reviewed the literature, intent on reproducing change-detection algorithms used in published articles but needed the code (which was never included with the articles). He soon found out that even when he obtained the code, it represented only one step in many that would be needed to reproduce the algorithm.

Adar's frustration grew as he tried to evaluate and compare different approaches. He needed not only the code but a multitude of other information to get it up and running. The list is long but includes things like the exact version of the computing language, operating system, dependency files, original data sets, packages … important things that weren't currently part of the curation process.

Adar developed a four-step process that, in some cases, took months of work to evaluate one single algorithm from a published paper.

1. Code was the first step. Sometimes the code was available as *pseudocode* through the published papers. Sometimes it was available in one of the many repositories. Sometimes he needed to contact the authors. Adar was sometimes able to find code even when the papers didn't mention that there was any. And sometimes he had to write the routines himself. This took time.

2. The second step was to get the proper hardware. Some of the implementations required very extensive or particular hardware. This took time, too.
3. The third step was to reproduce the run environment. Some of the algorithms ran under Windows, others under Linux. Adar needed to obtain and install the correct operating system versions. The rest of the environment also had to be configured correctly, with the right libraries, dependencies, and parameters. This took a lot of time. "In many cases you have the software code, and it relies on another package. You find and install that package, and you find that *it* has a dependency." In one case, it took seven steps for Adar to find the bottom rung on the dependency ladder.
4. The fourth step was to locate and debug the other installation errors, a process that could take from several days to weeks.

The core issue is that we need to curate today differently than in the past due to the fact that technologically enabled research is forever evolving.

Adar found his next challenge and embarked on a two-year project to develop Code Ocean. At Cornell Tech's Jacobs Institute, Adar and colleagues developed a cloud-based *executable* environment that allowed researchers not only to deposit their code but all other dependencies in order for their scientific software to run. There is an emphasis to curate all information for reproducibility and not just the article. As our technology evolves with new programming language versions and upgraded operating systems, we need curation tools that capture the needed information, data, and systems so research can be preserved for the future.

Code Ocean's goal is to slash the time needed to implement third-party software by providing online run environments. For example, a researcher might accompany a paper with software that has been published to the Code Ocean platform. The code, source files, and input data are available, so that the reader can run the program to reproduce the original researcher's results … or, perhaps, change the parameters and input data to test additional scenarios.

Readers can access Code Ocean through a widget attached to the paper on a publisher's website. Or they can go directly to the company's own website to search and use the code archive as a resource, and the virtual computing environment as a development and collaboration tool.

As research and technology evolves, so should curation.

Data-Intensive Research Architectures

Eric Whyne, founder of DataMachines.io, and contractor building data-intensive research architectures for the Defense Advanced Research Projects Agency.

Among other DARPA projects, Whyne developed software to analyze very large data sets, including the XDATA project. These projects produced publications. And they produced code. And code developed with federal funding is supposed to be made available to the public as "government off-the-shelf" software, or GOTS. The trouble was that transferring the software got very cumbersome very fast.

After some study, Whyne and his colleagues decided to take an open-source approach, and designed the DARPA Open Catalog as a self-perpetuating archive. They rejected the idea of "having everybody bring their code to us," and created Open Catalog as a list of links to code

repositories. There was no review of the software (too time-intensive), and no bureaucracy or red tape connected with the release.

The first step in getting Open Catalog running was social, rather than technical: face-to-face meetings with developers to persuade them to consider ways of making their software permanently available. In most cases, that meant loading the code to their organizational code repositories or to GitHub, and supplying links.

The second step in the project was to establish a "sensible license structure." Publishing software does not necessarily give users rights to use it. The GNU General Public License (GPL), for example, is a "copyleft" license, which means that derivative works may only be distributed under the same license terms as the original work. Thus, products derived from open-source material must be distributed as open-source themselves. That is not a permissive license but a poison pill under federal acquisition regulations, Whyne observed. The letter of the law prohibits using GPL software as part of a government product. (GPL software is used frequently nonetheless.)

After review and some close consultation with the Apache Software Foundation, the DARPA group decided to recommend, though not mandate, the Apache Software License Version 2.0. [12]

At that point, DARPA Open Catalog had a long list of code repositories with unknown software licensing provisions. The project team developed a machine-learning application to search the repositories, and tag them with the proper license structure.

The Open Catalog team wanted to make sure that researchers could reuse the code. Quick-start guides and instructions on installation environments were required. As expected, the software was heterogeneous and ran in many computing environments. Given the degree of variability, automating the documentation review wasn't feasible, so it became a manual process.

Open Catalog … got a lot of coverage. Whyne's favorite comment from the period came from a Russian publication, which called the Open Catalog team 'pathologically pacifist' for publishing this mass of material openly.

Open Catalog is still operating, still being published and updated. It caught attention from publications like *Wired*, *Popular Science*, and *Endgadget*, and got a lot of coverage. Whyne's favorite comment from the period came from a Russian publication, which called the Open Catalog team "pathologically pacifist" for publishing this mass of material openly.

The natural sequel to publishing open code is to publish open data. Thanks to privacy concerns and other restrictions on how data may be distributed and used, though, publishing data is a knottier problem. In some cases, even data gathered from open sources (like social media) cannot be redistributed or republished.

Whyne's presentation made a distinction between *reproducibility* and *replication*: if one can apply the method (or algorithm or code) to the original data and produce results that are statistically or formally the same as the original publication's, then the experiment is *reproduced*. If one can apply the method (or algorithm or code) to *new* data gathered according to the protocol, and obtain results consistent with the original publication, then the experiment is *replicated*.

Reproduced results may be exactly the same as the original results. Replicated results may not be exactly the same, but will be completely compatible with the original.

"I design data-intensive research architectures," Whyne noted. "I build clouds for DARPA. I have around 1,000 users right now." The volume of data is huge, and it's hard to move it from system to system. We need to start treating infrastructure as code, he said. There are about seven viable commercial cloud providers — including Amazon Web Services, Google Compute, Softlayer, Digital Ocean, Atlantic — and many, many clouds run by public agencies, the National Science Foundation, the National Laboratories, and federally funded university systems. If we treat infrastructure as code, we will be able to reproduce and reuse software across these systems much more easily.

During Q&A, a participant said he was intrigued by the statement that infrastructure should be considered as code. Does that mean that infrastructure is a research object, like code, or does it mean translating some of our thinking about code to apply it to infrastructure?

Whyne answered that public cloud providers, and the infrastructure that people are able to build, are moving at a lightning pace. No capability developed now will be valid in as little as three to five years, unless it is actively developed. So the author can package the software and the environment, so that they start and run in this kernel, or the author can abstract the conditions into a code that represents the infrastructure. Interesting projects include a piece of software called Vagrant, which is just an amazing thing for development environment. It allows users to provide one file, a Vagrant file. When it is run, it provisions an entire infrastructure environment that emulates the author's development environment.

Some of the people running infrastructure environments are system administrators, not software developers. They're not used to working with code, and probably won't move to infrastructure as code.

Distributed configuration management solutions have gotten a lot better over the last few years. The first ones were two pieces of software called Chef and Puppet. They allowed users to take a series of computer hosts, and then project a configuration/environment package-management environment onto them. Whyne wouldn't advocate their use now. They are certainly still valid, but there are better tools, like Ansible, which has completely taken their place in Whyne's work.

With such code, a software developer can describe an infrastructure environment, and then universally replicate that environment and reproduce it across other systems.

There's a side effect: some of the people running infrastructure environments are system administrators, not software developers. They're not used to working with code, and probably won't move to infrastructure as code.

Google published a very interesting book called *Site Reliability Engineering*, which also talks about this. The large public cloud providers are already treating their infrastructure as code, too. That's how they scale, and that's the only way to scale. The book advises hiring code writers to run infrastructure, because coders won't tolerate manual interactions with the systems. They will automate things, and they will abstract things, and they will make changes idempotent: if you

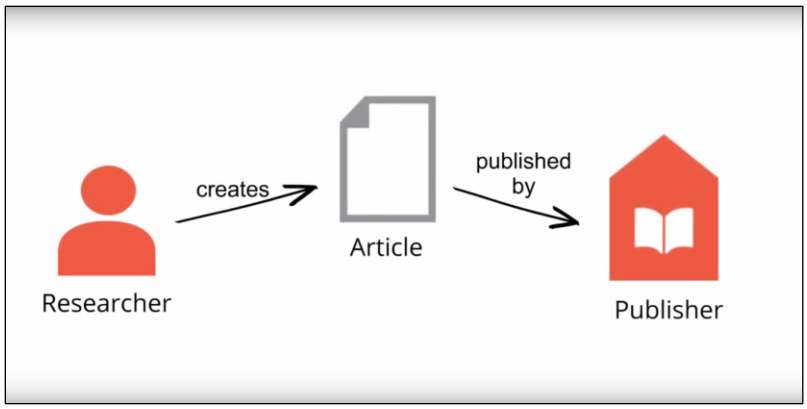
implement the same change twice, you're not going to get an inconsistent state, you're going to be able to project the same state. There are lots of nuances to running systems in this manner that haven't been widely spread yet. They need to get out there. We need to treat infrastructure as code.

Content, Context, and Curation

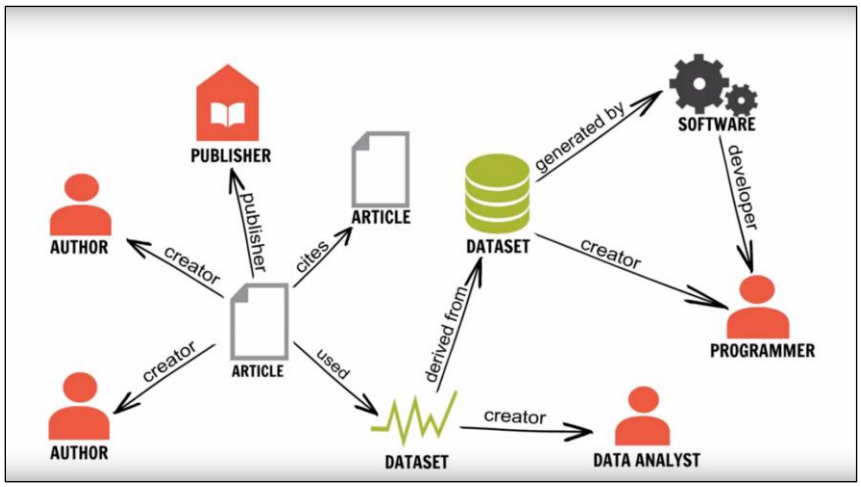
Sheila Morrissey, senior researcher at Ithaka, where she has worked with Portico, a service for preserving the digital artifacts of scholarly communications.

Morrissey introduced herself as "somebody professionally interested not in the future, but in the past, not what's next, but what was." She noted a convergence between current emphases on reproducibility and the methods evolved over 20 years of preserving traditional academic literature. When Portico got started, researchers did their research and produced a scholarly article that a publisher published. The printed article was the sole artifact of scholarly communication.

Traditional Publishing vs. Today's Publishing



Traditional publishing



Modern publishing with multiple artifacts (from the RMap Project, http://rmap-project.info/rmap/?page_id=98))

During the transition to digital preservation, the focus shifted from the print object to a digital analog of the print object. We are learning that digital content is complex. Now, a publication can include many artifacts, related and nested, created at different times in different places by different people. And these elements can continue to change as they are updated, expanded, corrected, revised. Tracking and preserving different versions is as important a part of the preservation process as it is of the research evaluation and review process, since it is critical to know exactly which state of the evolving work was the focus of which review or test.

Using these artifacts is different from picking up a print article or downloading a PDF. There may be original code, data, videos, animations, simulations, and other elements that require specific software. All of this makes the work more complexly mediated for the reader. In this rapidly changing environment, we must also maintain the flexibility to provide for unanticipated uses of existing content, and the addition of unimagined new kinds of content in the future.

A well-designed, large-scale taxonomy, for example, might tie related content together across multiple platforms and data types. Scholarly communication becomes not an object, not a journal article, but a network of heterodox objects, each consisting of multiple constituents, all of which can exist in multiple versions, and all complexly joined to the others. This presents obvious challenges to preserving, rendering, and reusing the "publication" and its underlying data and tools. As an example, Morrissey reviewed what was required for two New Zealand researchers to reproduce what is probably the world's first digital music, Alan Turing's Manchester Electronic Computer Mark II playing *God Save the Queen*, *Baa Baa Black Sheep*, and *In the Mood* in 1951 (sidebar).

Scholarly communication becomes not an object, not a journal article, but a network of heterodox objects, each consisting of multiple constituents, all of which can exist in multiple versions, and all complexly joined to the others.

We have 20 years of experience in identifying and preserving the objects necessary for verifying reproducibility, and with the kinds of infrastructure necessary to ensure that these objects are discoverable, accessible, and usable. "There are models, there are vocabularies, and there are formalisms of expression, ontologies, and languages that have been developed for preservation and that are directly applicable to the problems of ensuring the reproducibility of scientific results and curating those outputs," Morrissey said.

The field has developed pragmatic mechanisms for acquiring content, managing data on a large scale, and enduringly identifying objects to arbitrary levels of granularity. In the works are repositories being developed with these requirements in mind.

"Emulation as service" is evolving. Some issues remain, but it is much closer to practice than it was when it was conceived 20 years ago. And we are developing intuitions about what is needed to emulate an original experience, and what the limits of such emulation are.

Efforts to automate the capture and assembly of artifacts and context are beginning. That's vital: experience suggests that if it can't be automated, it won't happen. Who has the time to capture and categorize all of this material manually?

Preservation has somewhat different time horizons and concerns, but there is substantial common ground, in the objects concerned and the long-term objectives of research curation and reproducibility.

Sidebar: Reproducing the First Digital Music

In 2016, two New Zealand researchers, Jack Copeland and Jason Long, reproduced the first digital music: Alan Turing's Manchester Electronic Computer Mark II playing *God Save the Queen*, *Baa Baa Black Sheep*, and *In the Mood* in 1951.

Turing's computers included an audible signaling function, which he called "The Hooter." It was a familiar device that emitted a click when triggered. Turing and company soon discovered that repeating the click at a fixed interval created a tone, and periodically changing the interval could produce a series of notes of pitch and duration. Turing used these notes to signal what the computer was doing. A schoolteacher and musician (and eventually one of Britain's most distinguished computer scientists) named Christopher Strachey got hold of the *Programmers' Handbook for Manchester Electronic Computer Mark II,* and started trying to make music.

In short order, Copeland and Long wrote, "Strachey turned up at Turing's Manchester lab with what was at the time the longest computer program ever to be attempted." After a night-long programming session, Strachey's first, the machine "raucously hooted out the National Anthem." Strachey expanded the Mark II's repertoire, and eventually a crew from the BBC arrived to record a short medley of tunes. They cut an acetate disk using a mobile recorder whose turntable spun too fast, with some wow and flutter tossed in.

The original computer programs are long gone. But with a limited set of artifacts — Turing's description of how the Hooter worked, the programming manual, a handbook for the mobile record-cutter, and an audiotape made from the unreliable BBC disk — Copeland and Long recreated routines that would play the tunes on the record as they would have sounded emerging from the Hooter's speaker. They began with the audio recording, focusing on tones that the Hooter could not have produced. This gave them a roadmap for processing the entire tape to counter frailties of the mobile acetate recorder. Once they knew the notes to sing, they could use the Mark II manual's instruction set to "disassemble" the audio output into Mark II Hooter code.

This is a textbook example of the difficulty of reproducing computer science results, and it underscores the importance of a contextual network. The reproduction is itself a research project. But the story may not end cleanly when reproducibility is demonstrated. "How do we evaluate that claim of fidelity?" Morrissey said. "How do we make an assessment of this particular research project?"

Group Reports on New Forms of Content and Radically New Approaches to Content Creation

Time-honored practices in publishing are currently undergoing tumultuous disruption on a number of fronts. After the appearance of digital archives of downloadable PDF versions of published articles, the proliferation of new research-sharing models grew quickly to include increasingly sophisticated self-archiving, preprint servers, and early university research repositories. The landscape has continued to change with increasingly popular web-based collaboration tools that support not only collaborative writing, but also code development and data curation.

Critical elements; curated updates; finding and indexing (Report from Group A)

The questions for discussion:

- *What new content is most critical to reproducibility: raw data, algorithms, code?*
- *For code there are typically updates and fixes. Will curated updates occur for all forms of content and information for maximum reproducibility?*
- *How will new content get found and indexed?*
- *Other — please identify any other items we should consider in the future.*

There is no single most important component to reproducibility. All of them are of equal importance and the research cannot be reproduced without all of the ingredients.

"Raw data," however, may not be sufficient, even with the appropriate code and algorithms. There must also be clear descriptions of how the data were collected and cleaned up; this could be part of the metadata associated with the "data" research object.

A key issue is that each of the research objects associated with a publication (or an experiment) can live its own life. As soon as possible, publishers need to standardize their metadata to assign digital object identifiers (DOIs) to code, software, data, and other objects in order to make them discoverable.

It is particularly important that publishers and authors have faith in the sustainability of any external repository, satisfying the primary condition of preservation. Standards do exist, but they must be adopted by all publishers, and publishers need to start this discussion in earnest as soon as possible.

Right now, it is more important to develop methods and standards for maintaining software than for maintaining data. There are already standards for archiving data; there don't seem to be any yet for archiving software.

Until these standards are available, publishers themselves should hold software in "dark archives." In these, the software might not be properly searchable or generally available, but it would be preserved against the time when standards emerge.

There is room for discussion about what the interim archive would look like. Each publisher might build its own, or it might be done in a coalition, much as CrossRef was established to standardize indexing. Publishers should meet to start the process of creating broader metadata in the very near future.

Metadata should clearly identify authors and contributors of individual research objects. Not all of the authors of a primary paper should receive credit for each of the associated software and data objects. This is fundamental if these elements are to become independent, citable scholarly objects.

Dealing with new content; tools for collaboration (Report from Group B)

The questions for discussion:

- *New forms of content will spring up in unanticipated ways from the researchers themselves. Should we try to put methodologies in place that could deal with any type of content, or try to predict new forms of content and adapt methodologies to each specifically (or a combination of both)?*
- *Are there collaborative tools to support data curation, and are there lists or web links to these collaborative tools to help with data curation? (Example: https://www.whitehouse.gov/blog/2016/10/28/federally-funded-research-results-are-becoming-more-open-and-accessible)*
- *Other — please identify any other items we should consider in the future.*

Publishers, research communities, and funders will need to collaborate to develop approaches to managing new forms of content. It cannot be left to any single stakeholder group. Publishers may wish to consider coordinating this development as a new line of business.

There are existing models for dealing with new kinds of content. In the social sciences, for example, ICPSR (the Inter-University Consortium for Political and Social Research at the University of Michigan) offers methods for publishing data in the social sciences. In bioinformatics, GenBank, part of the National Center for Biotechnology Information at the National Institutes of Health (NIH), is a 30-year-old open resource of genetic data and metadata linked with publications.

Timeliness is a concern. How long should archivers plan to ensure that the contributed code will run on the specified platforms? All the elements of computing infrastructure are changing constantly. A repository may be able to provide encapsulated environments that replicate part of the original environment, but even these may be operational for only a fixed time.

It is possible to envision an evolution of approaches to archiving, and of objects themselves, in which the important or successful approaches and objects will survive and the less important and less successful will become extinct.

To narrow the problem, it might be fruitful to focus on content in areas significant to IEEE and its members, though that itself covers a wide area. This experience might teach us enough to allow archivers to begin to predict what may arise in the future, and begin development that is more anticipatory than reactive.

Another possibility: Develop a research equivalent of the U.S. Library of Congress, rather than individual repositories for each community or discipline. There was also discussion about whether there should be an equivalent of the Libraries of Congress rather than just one for each different community. The analogy to the Library of Congress also raised the question about whether this effort should be centered on the U.S., or international in scope. Since the research is a worldwide enterprise, as are ACM and IEEE, the wider approach makes more sense.

A number of collaborative tools already exist, though it is unclear how many will survive in the long run. Google Code is gone. SourceForge is still operating, but not the force it once was. GitHub has emerged as a dominant tool — perhaps unexpectedly. A workshop on this topic five years ago would not have anticipated its growth. *Git,* with its version snapshots and hashed signatures, is an elegant tool for collaboration. A similar approach to archiving data would be an important development, but it doesn't seem to be in the works now.

Cross-platform integration makes collaboration possible through application program interfaces (APIs), without forcing Borg-like assimilation. Site availability is an issue. Uptime varies from resource to resource, and portions of the collaboration or archiving service black out when a component web site goes down.

Different communities produce different data. Should there be metadata formats for "commonly used" data types? This is an open question.

If data is to be published along with other kinds of artifacts, how will that data be validated for correctness, completeness, and usefulness? Little data validation is being done today, even by the funding agencies that mandate data management plans and publication of data. NSF does some minimal review of published work, but it is not a real validation.

Cross-platform integration makes collaboration possible through APIs, without forcing Borg-like assimilation.

Other questions include, "Should data, too, be individually peer-reviewed?" and "How do we reconcile the different timeframes expected for publishing data in different disciplines?"

There will be a need to standardize. We'll need to standardize the data itself — agreeing on what we keep, in what format, and how we ensure that it's complete. A more difficult challenge is to standardize the research process itself and the documentation for that research to promote reproducibility in the future.

Essential products, practical actions (Report from Group C)

The questions for discussion:

- *What are the essential products of scholarly engineering research? How will these be likely to change in the future?*
- *What do you think you and the other participants here can do to advance the reproducibility of research after this workshop is over?*
- *Other — please identify any other items we should consider in the future.*

The first issue was defining "reproducibility." Different research communities define reproducibility differently, and will have different attitudes. And because of this, it may not be feasible to develop a uniform standard for what constitutes reproducible research. It might be preferable to develop approaches to help the individual communities adopt their own definitions and practices, reflecting their own priorities.

As for specific actions that workshop participants can take: The participants can capture as much of the process as possible, identifying what's needed, what the social and technical aspects are,

how different communities define reproducibility and the associated concepts. It might be useful to identify several particularly diversified research communities (whose research pulls technical elements and norms from a number of disciplines), and then to support pilot programs promoting reproducibility in those communities. This activity would also reveal what reproducibility efforts are already underway in those communities. These existing efforts could be the seeds of new pilot programs.

Pilots in these diverse communities might yield insights on how to approach the greater diversity of science as a whole. If the pilots are successful, they should attract more participants and spark more pilots. These could show the way to propagate reproducibility.

Workshop participants can also help make reproducible research more visible and influential in their own communities, making the benefits broadly clear. Thus, an experimental result is not merely right or wrong. By being reproducible and offering reusable artifacts, the research is a springboard to further science.

One group member suggested that setting goals in a high-profile initiative might help increase awareness of reproducible research's benefits. The United Nations' Millennium Goals for sustainable development are an example (though a wider and much more ambitious example). Setting up a challenge with a deadline helps focus energy and attention. This is something that publishers, funding agencies, and societies could collaborate on.

Overall, the most productive approach should be to encourage and reward reproducible research, rather than try to coerce it. The issue of funding arises right after the issue of defining reproducibility. The funding agencies need to continue to build more support for, and incentives to do, reproducible research. One participant stressed that the funding issue is "massively consequential." If we are going to insist on improving the reproducibility situation greatly, then it will add to the cost of the research. Surely, many researchers would love to do very high-quality, highly reproducible work, and then they look at the grant they get from NSF. "You do what you can with the money you've got. There's a downstream consequence."

Curatorial challenges; community advocacy (Report from Group D)

The questions for discussion:

- *What are the most significant curatorial and preservation challenges presented by complex, distributed digital artifacts of scholarly communication?*
- *Should we have a community group that would advocate for forms of content that is more easily reproducible or would set industry standards for content?*
- *Other — please identify any other items we should consider in the future.*

The workgroup asked two questions: What are the most significant commercial and preservation challenges presented by complex distributed digital artifacts of scholarly communication? But to whom are the challenges presented? Stakeholders include the funders and the publishing industry (the publishers themselves and all the ecosystems around them), the researchers, research institutes, the policy makers, readers, libraries, and the technology providers. Challenges were considered for each respective entity. Some data, for example, might require specific hardware or software to open. Merely mandating that it be deposited may not accomplish the purpose of promoting reproducibility. Each area needs to be examined to see when mandated data deposits make sense.

Carrot-and-stick incentives from funders are a good framework, but funding organizations need help to define the carrots and the sticks. Making policies with the wrong incentive structures risks "build it and they will *not* come." We agree that communities and funders need to work out these structures together.

Among other challenges, the group considered metadata, the challenges of deciding what information is necessary, and of gathering and maintaining it. It appeared that it might not be obvious what kinds of new information need to be attached to data and code. Versioning is an issue, controlling the processes of creating, uploading, updating, and forking (creating new version lines), so that all of the objects of evolving code or data that correspond to a publication will be available, bugs and all.

Durability of services is another challenge. For example, Code Ocean is a start-up company, providing a service of replicating computing environments in the cloud. What happens a few years from now? What happens with GitHub? Google Code is no longer available. So it's not just about having appropriate reproducibility services available, it's also about developing processes to assure that content will be preserved even if the original preserver no longer exists.

Researchers face many challenges. There are intellectual property issues. There's the discomfort of letting peers see rough code. There's the need to exploit their data for as many publications as possible — and how much exclusivity the funders will permit. We need to modify the way we measure the impact of research projects.

It's not just about having appropriate reproducibility services available, it's also about developing processes to assure that content will be preserved even if the original preserver no longer exists.

Research institutes include many smaller sub-entities, including repositories and tenure committees. How does one measure scholarly productivity aimed at reproducibility, and how can it be made consistent with policies in the rest of the university, and with other universities (allowing for the way scholars move among institutions)? How can this productivity be reported to funders and policy-makers?

One may think at first that "readers" are other researchers, but they are also the public at large. Is the general public also entitled to access research artifacts? If somebody has cancer, and is paying taxes to support the research, should she or he be able to read the research and use the data? Do we facilitate this, or not? Companies use the published research and artifacts to increase their pace of innovation. Are these companies then stakeholders in reproducible research?

Should we invest in the technology to be able to reproduce the research reported in each and every article 50, 100, or 500 years from now? Should there be some kind of a limit, saying that this research will be reproducible for this number of years, and that after that it will just cost too much maintain the ability to reproduce this research?

How is the research community to be educated? What exactly should we tell researchers to use as the gold standard of reproducibility, because there isn't a gold standard yet? And should there be a community group that would advocate for forms of content that are more easily reproducible, or would set industry standards for content? The answer to this last question is, "Yes." This workshop is the answer. Because this is the first workshop on the topic (in the engineering and

computational science community), it is very important to gather stakeholders from all the different entities to have those discussions, create consensus, and overcome all of the challenges.

One participant addressed a "very intriguing idea about incentives." Take, for example, a curated data set that has taken a long time to put together. Right now, the incentive is for the researcher to hold the data back, to get the maximum number of publications out of it, because that's where the recognition comes from. If the data set can become an independent object with its own DOI, an object that's citable, downloadable, and linkable, this becomes a way for the researcher to get recognition for that object.

Adar concurred that there is a need to more formally recognize and define those kinds of artifacts as first-class research outputs. Maybe that is something that can be done on a policy or regulatory or funding or other top-down basis.

Baillieul noted that the presentations had subtly raised the idea that making the transition to treating software and data objects as primary research products needs to be done with small and enthusiastic communities that may get some additional portion of their funds for the research needed to develop these things as prototypes and models. At present, stakeholders are a long way from having very large-scale repositories that are going to work across the broad spectrum of research communities.

Right now, the incentive is for the researcher to hold the data back, to get the maximum number of publications out of it, because that's where the recognition comes from. If the data set can become an independent object with its own DOI, an object that's citable, downloadable, and linkable, this becomes a way for the researcher to get recognition for that object.

A participant cited ImageNet, a very large set of images that researchers have been using for experiments. Would NSF consider that as a first-class object? The answer is, "Yes." NSF does acknowledge data sets and those kinds of products, both in the bio sketch since 2013 and in the annual and final reports. Whether such contributions would be evaluated by a merit review panel or a tenure committee is not under NSF control.

Plenary Panel: Peer Review and Quality Assurance

As nontraditional types of research products (i.e., data and software) become a significant component of the curated research record, how should quality assurance be organized? Some questions to be pondered: Do we need to provide a common platform? Can we run experiments using different software and environments? How to address the possibility of proprietary software (e.g., compilers)?

Panel moderator: Sheila Hemami, Director, Strategic Technical Opportunities, Draper Laboratory.

Panelists: Bernie Rous, ACM; Jennifer Turgeon, Sandia National Labs; Eleonora Presani, Product Manager, Scopus, Elsevier.

Data, Software, and Reproducibility in Publication

Bernie Rous, Director of Publications (now emeritus) at the Association for Computing Machinery, and chairman of CrossRef's Board of Directors.

Rous began by observing that the workshop discussions indicated that some consensus is forming about the complex issues of reproducibility. For publishers, the question is, "What services and tools will publishers need to develop or to provide for the authors, for reviewers, and for end users to support the publication of data and software artifacts, and to facilitate their reuse."

In the future, publication of experimental research articles (as distinguished from formal proofs on the one hand, and descriptions of phenomena on the other) must, at very least, include access to the artifacts necessary to reproduce the reported results. It's important to note that "reproducibility" has no single technical definition. "Reproducibility" can mean different things, and scientists use many other terms in an attempt to specify levels or types of reproducibility. And some of these same terms are used interchangeably: repeatability, replicability, verifiability, reusability, and validation are a few of these.

In computer science, though, there is no tradition of affirming reproducibility in publication.

In some disciplines, reproducibility of results is a traditional expectation for both ethical and legal reasons. In computer science, though, there is no tradition of affirming reproducibility in publication. The reasons for this are instructive. Set aside the tremendous variability of instrumentation, data sets, software, and computational resources. Computer scientists often study the hardware and software themselves, rather than natural phenomena.

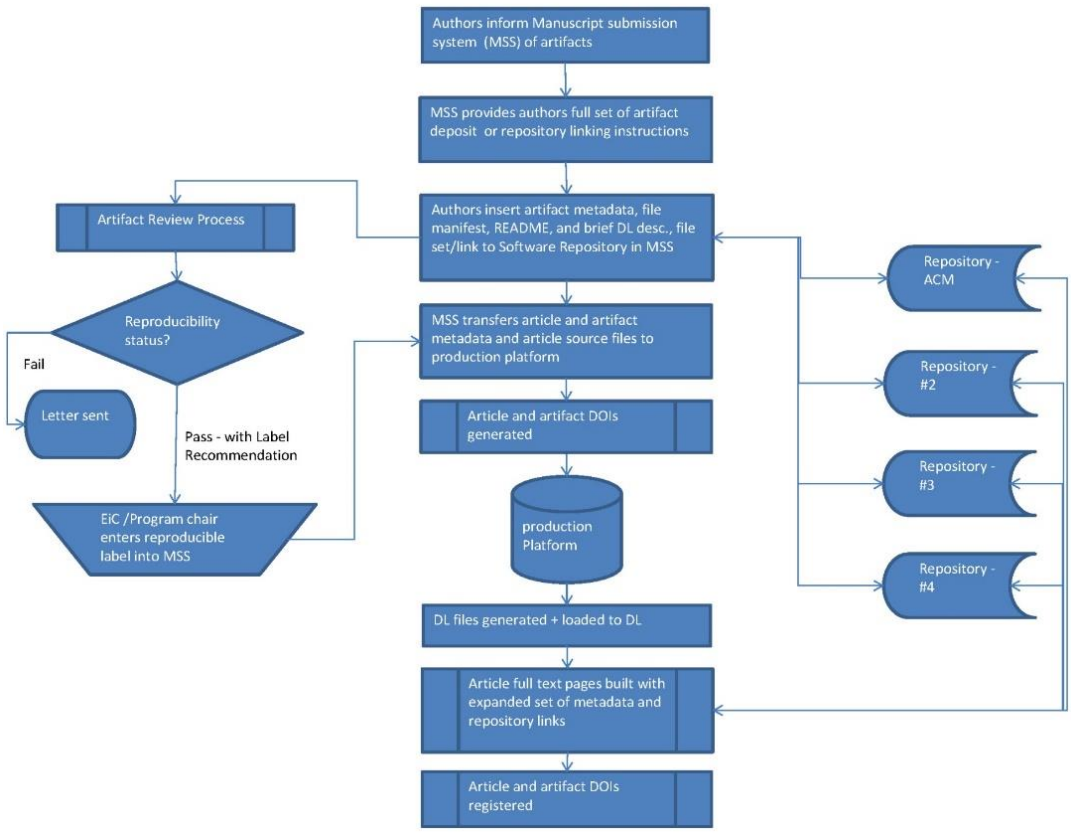
Reproducibility is an issue even before publication, during the review process. ACM publishes 500 volumes of conference proceedings each year. More than 15 ACM-sponsored conferences have already undertaken artifact review, that is, review of the data and software submitted along with the conference papers.

At the simplest level, it has proven extremely difficult to recreate the environment for running the experiment. Authors must provide extensive documentation before any attempt can begin, and even so, reviewers commonly spend weeks trying to recreate the experimental environment …

only to fail in the end or give up under the constraints of time and the limits of frustration: the author may have been using a different version of an operating system, or failed to mention libraries or routines, or forgotten to specify some variables, or left out some key scripts.

At another level, the reviewer may not have access to the same computational resources. This happens often in experiments with high-performance computing.

Two years ago, the Association for Computing Machinery assembled a task force on data, software, reproducibility, and publication. The 35 members included members of ACM Special Interest Groups and journals who had initiated reproducibility reviews, as well as representatives of IEEE and the Society for Industrial and Applied Mathematics (SIAM), and others, as well as ACM.



An ideal publishing workflow supporting peer review, artifact review, reproducibility verification, and integration of objects. (B. Rous and C. Rodkin, ACM)

Two things became apparent at the first meeting: first, each of the groups was trying to solve similar problems, and each was trying to solve the problems alone. The task force was the first step in getting these stakeholders to seek to define the problems common to their different areas, and to begin coordinating efforts to find a common solution. Second, the existing siloed efforts were entirely divorced from the publication process.

The ACM task force effort has produced some concrete results, Rous said. First,

> "…we learned that this is really very, very early days in terms of reaching a single, general publishing solution that enables ubiquitous reproducibility of experimental results. We're very far away from some imagined holy grail where machine-readable metadata describes an experiment and its methodology in such accurate detail that another machine could read that metadata, automatically assemble the environment, pull in the data sets and launch the software successfully. What we need to do really is start with very, very little steps."

Second, it's clear that publishers can play a role by building tools to support the objects needed to confirm reproducibility, and by creating incentives for researchers to change their habits and submit these objects. The task force is drafting best-practices guidelines for data, software, and reproducibility reviews in publication. It should be released in 2017.

Third, publishers should *not* require confirmation of reproducibility as a criterion for acceptance of the article. That is too high a bar, and it would impose unacceptable additional loads on a peer-review system that is already overstressed. The intensive labor needed to reproduce an experiment would be crippling if it were mandatory. Instead, paper review and acceptance should be independent of the review of the artifact.

Fourth, it may well be that different reviewers are better equipped to handle artifact review and paper review. Different expertise may be needed in each case. The subject domain expert is not always the most adept at installing and rerunning a piece of software.

Fifth, anonymous review is *not* necessarily desirable for data and software — at least initially, and possible not ever. Experience has shown that setting up and running an experiment correctly requires considerable communication between author and reviewer.

Publishers should not require confirmation of reproducibility as a criterion for acceptance of the article. That is too high a bar.

Sixth, it is important to motivate authors and reviewers to support the reproducibility process. Authors must commit to substantial extra work to prepare their artifacts for review. To motivate and recognize reproducible research, ACM has developed a system of badges that flag papers that have passed different levels of reproducibility review. Papers that do not pass reproducibility review are not stigmatized. The incentives are all carrots, and no sticks.

Reviewing artifacts requires either specialized artifact review panels, or publications need to give the reviewers public credit … which is possible if artifact reviews are not anonymous. Artifact reviewers can be acknowledged in the papers they review. In ACM's conference publications and in three of the association's journals, artifact reviewers may publish short companion papers that describe in detail what they discovered in setting up and rerunning the experiment. Thus, the reviewers get credit for a publication, and they may also get a position on the masthead as algorithms editor, artifact reviewer, or a similar title.

In June 2015, *ACM Transactions on Mathematical Software* launched a replicated computational results initiative (continuing the journal's 40-year tradition of publishing software associated with about a third of its papers). The first paper in that issue, by F.G. Van Zee and R.A. van de Geijn, was subjected to artifact review.[13] It passed, and the first page of the article carried

the blue "Results Replicated" badge, an editorial note that the computational results have been replicated, and a link to the second article in the issue, J.M. Willenbring's report on the artifact review. The review report includes its own source material (including a video), a description of the process, and a link back to the Van Zee and van de Geijn publication. Both the paper and the artifact review include source materials.

It's very important to develop reproducibility labels with clear definitions. ACM has evolved a taxonomy of standard badges, differentiating levels of artifact review: from "Artifacts Available" through "Artifacts Evaluated — Functional," "Artifacts Available — Reusable," and "Results Replicated" to "Results Reproduced."

ACM Reproducibility Badging

Small icons for search results and ToCs, large for PDFs.

Since the terminology of reproducibility is not yet standardized, each badge links to a definition. The badges represent levels of reproducibility review that can be added to the paper's metadata.

The badges, links to their definitions, and links to the artifacts, are all currently inserted manually. This kind of piecework curation does not scale, however. The process should be automated, and to do that requires standard definitions and metadata descriptions for data sets and software, so that data and software can become first-class objects in their own right and be identified, searched, cited, and linked. As first-class objects, the data and software merit their own citation papers; they would no longer be relegated to "supplemental material" available only through the paper.

ACM has completed an internal draft of an XML schema for artifact metadata, which is now ready for wider distribution and comment. Rous expects that other bibliographic metadata schemas, such as CrossRef's (crossref.org) will evolve to accommodate extended metadata deposits with independent DOI assignments.

It's also important to realize that artifact review and reproducibility badging of the primary paper is independent of publishing the artifacts themselves. This is the reason for the orthogonal "Artifacts Available" badge.

Even without the artifact, the reproducibility badge indicates that the paper's results are trustworthy. The cachet of the badge is also an incentive for other researchers to participate in

artifact review. And it is also possible that some authors might make their artifacts available without participating in the optional reproducibility review.

It is necessary to recognize, however, that proprietary interests may prevent researchers from making their software and data available for review. Corporate researchers publish, companies fund academic research, and more than a few academic researchers exploit their research to found a commercial venture or drive a continuing series of publications. They may, on the other hand, be willing to submit these artifacts for confidential review, if not for publication.

When an author does grant permission for publication, the publisher must be ready to serve the artifacts to the readers, out of its own digital library and accompanied by a thorough readme file that will help the user set up and rerun the experiment. Publishers must be equally prepared to link to artifacts preserved in external repositories, GitHub, for example, while preserving the related metadata, including reproducibility badging.

With so many software curation platforms emerging and evolving, it is too early to tell which, if any, will be self-sustaining in the long run. If there are any doubts at all about the longevity of external repositories, publishers should encourage authors to submit, at very least, a back-up zip file containing as much of the artifact or artifacts as possible, with installation and operation instructions. Publishers should be prepared to serve these packages from their own sites, if necessary.

The community also needs to develop the legal framework for serving artifacts. Software and data fall under different IP regimes, but in either case, the publisher needs to specify ownership, the user's rights, and the terms of liability. These should be established and managed independently of rights to the paper. ACM is in the process of drafting such a legal framework. When legal review and revision are done, ACM will deploy it in its e-rights management application.

As noted several times, one of the knottiest aspects of rerunning experiments can be recreating the computing environment.

Publishers must be prepared to link to artifacts preserved in external repositories, GitHub, for example, while preserving the related metadata, including reproducibility badging.

Developers are creating lightweight virtual machines and wrappers that provide the environment needed to run the software. These encapsulation tools are, so far, unstandardized and vary from service to service. Rous suggested that authors might benefit from template instructions for building such wrappers. Under a small Sloan Foundation grant, ACM is running three pilot programs, experimenting with methods for integrating its Digital Library content with three different external software platforms. A major aim of the project is to study the encapsulation process to extract generalized instructions and templates to guide authors in preparing objects for deposit in each of the platforms.

To some extent, creating these wrappers parallels the work the author must do to write a readme file that will allow end users to recreate their experiments in their own systems. In creating a wrapper, though, the author builds this template earlier in the process, to facilitate the object review as well as end-user reproduction. Publishers are unlikely to provide virtual machines and computational resources themselves, so it is imperative that they integrate their offerings with the

specialized curation platforms and data repositories that do provide these features and functions. Ideally, the publisher will integrate so tightly with external repositories that the presentation appears seamless to the reader: the reader won't lose the context of the article on the publisher's platform when they access the artifacts to rerun the experiment on another.

Thus, a reader could launch a simulation from within an online article. The simulation runs on a third-party service. The user might have the ability to test the simulation with new data and parameters to generate new results. The user might even be able to edit and modify an image of the code in the external virtual machine. This will raise complicated publishing questions, of course. How should we capture the new user-generated result or the derivative artifact? How should we label its provenance and its relationship to the original work? Will it be considered reviewed or unreviewed? Will it be considered reproducible by inheritance? How will credit be given, and how will it be cited? This all remains to be worked out.

Software Quality Engineering: Paving the Path toward Research Reproducibility

Jennifer Turgeon, principal member of the technical staff at Sandia National Laboratories.

Sandia National Laboratories' role includes maintaining the U.S. nuclear arsenal, developing tools for national defense and national security, and studying climate, cybersecurity, high-performance computing, power production, and sustainable transportation, among other missions. It produces tremendous volumes of data and software, much of them part of programs that could present significant risks. Thus, a risk-based quality-control approach is an essential part of Sandia's mission. The quality objective is human and public safety, rather than publication.

The system depends ultimately on the researchers' integrity. Peer review is a cornerstone of the process. As an institution, Sandia wants its research to be both trustworthy and reproducible. Maintaining the integrity of Sandia's reputation is essential.

Across Sandia, some projects have small budgets. Some have massive budgets. One size emphatically does *not* fit all. It's important that quality assurance approaches be appropriate to both the resources and the potential risk, while still maintaining practices that support trustworthy and reproducible products. Using a risk-based approach, projects must determine what might happen if a software product fails. The likelihood and consequences surrounding the potential failure determine the level to which quality practices are utilized on the project. Sandia's basic quality standards, the expertise of the team, schedules and deliverables, and, of course, funding all go into the decision on the quality approach.

A Sandia research group focused on simulation and computing, which also publishes a significant amount of its research, has developed a set of 30 practices for ensuring that software and data are reliable and reproducible. The practices are scaled, based upon the risk level of the work being developed. The demands and resources of the project determine which practices are emphasized, and to what level of rigor. Constants include configuration management, backup and disaster recovery, testing, and release strategy.

Sandia National Laboratories Advanced Simulation and Computing Software Quality Engineering (SQE) Process
SQE Categories/Process Areas/Practices
Project Management SQE Category
1. Integrated Teaming
PR1. Document and maintain a strategic plan.
2. Graded Level of Formality
PR2. Perform a risk-based assessment, determine level of formality and applicable practices, and obtain approvals.
3. Measurement and Analysis
PR3. Document, monitor, and control lifecycle processes and their interdependencies, and obtain approvals.
PR4. Define, collect, and monitor appropriate process metrics.
PR5. Periodically evaluate quality issues and implement process improvements.
4. Requirements Development and Management
PR6. Identify stakeholders and other requirements sources.
PR7. Gather and manage stakeholders' expectations, requirements, and constraints.
PR8. Derive, negotiate, manage, and trace requirements.
5. Risk Management
PR9. Identify and analyze risk events.
PR10. Define, monitor, and implement the risk response.
6. Project Planning and Oversight
PR11. Create and manage the project plan.
PR12. Track project performance versus project plan and implement needed (i.e., corrective) actions.
Software Engineering SQE Category
7. Technical Solution
PR13. Communicate and review design.
PR14. Create required software and product documentation.
PR15. Identify and track third party software products and follow applicable agreements.
PR16. Identify, accept ownership, and manage assimilation of other software products.
8. Configuration Management
PR17. Perform version control of identified software product artifacts.
PR18. Record and track issues associated with the software product.
PR19. Ensure backup and disaster recovery of software product artifacts.
9. Product Integration
PR20. Plan and generate the release package.
PR21. Certify that the software product (code and its related artifacts) is ready for release and distribution.
10. Deployment and Lifecycle Support
PR22. Distribute release to customers.
PR23. Define and implement a customer support plan.
PR24. Implement the training identified in the customer support plan.
PR25. Evaluate customer feedback to determine customer satisfaction.
Software Verification SQE Category
11. Software Verification
PR26. Develop and maintain a software verification plan.
PR27. Conduct tests to demonstrate that acceptance criteria are met and to ensure that previously tested capabilities continue to perform as expected.
PR28. Conduct independent technical reviews to evaluate adequacy with respect to requirements.
Training Support Category
12. Training
PR29. Determine project team training needed to fulfill assigned roles and responsibilities.
PR30. Track training undertaken by project team.

The release strategy, coupled with other practices, is tied to reproducibility. When software or data are reproduced, regenerated, or shelved, that break-point constitutes a release, which must be packaged in such a way that the work can be rebuilt, if needed. Clearly, a project may have many releases along the way, and Sandia aims to ensure that each release can be reproduced through stringent version control. Verification, validation, and peer review — often external peer review — are important parts of the process. These, with requirements management and design discipline, are the core elements of every project.

Traditionally, software developers have treated a "release" as a packaged software product that can be reproduced or rebuilt at any given time. Input and output data have only recently become part of the release strategy, which is still being developed. Data had been seen as a byproduct of research, to be used, vetted, and peer-reviewed. But then, it would be left to evaporate. In some cases, little thought had been given to retaining and reusing it. This is an issue, though, that many engineering and computer science researchers must address. Sandia, too, must deal with versioning of data sets, as raw data becomes filtered data becomes abstracted data. It's vital to know which stage of the data one is dealing with and the design decisions and processes used to create it. Are we capturing the rationale behind the data manipulation, so that the same or similar output can be reproduced?

There's another issue being investigated: third-party software. Software that Sandia develops in-house typically is not a problem because existing procedures track releases and manage versions. But sometimes researchers bring in third-party software. That changes, too. The quality-control system hasn't thought through the implications of version changes in outside code, and we haven't developed ways to formally vet third-party software before integrating it with our products. Currently, projects are responsible for establishing a level of trust in the third-party software and define how the software will be managed within their own work.

The software development practices we have in place and continue to improve upon allow us to inherently develop reproducible software and data, using a graded approach to the level of rigor we place on those practices. Sandia strives to conduct both research and product development work in the same manner in order to establish and maintain the integrity we have earned through our past, current, and future work.

Quality Assurance for Nontraditional Research Products: Is There Any?

Eleonora Presani, product manager for Scopus, Elsevier's science abstract and indexing service.

Even today, the caches of published science hold a lot of data, software, protocols, code, and algorithms — but these treasures (and their authors) are mostly hidden away. The objects themselves are difficult to find, and it is unclear where they might be stored so that potential users can go looking for them. A major part of today's research output is thus buried in supplementary material, and is not easily credited or discovered. So their authors often don't get recognition for the work they've done to create these resources, or for their contributions to the formal publication. These obscurities and inefficiencies put a brake on the progress of science, and are a roadblock to reproducibility.

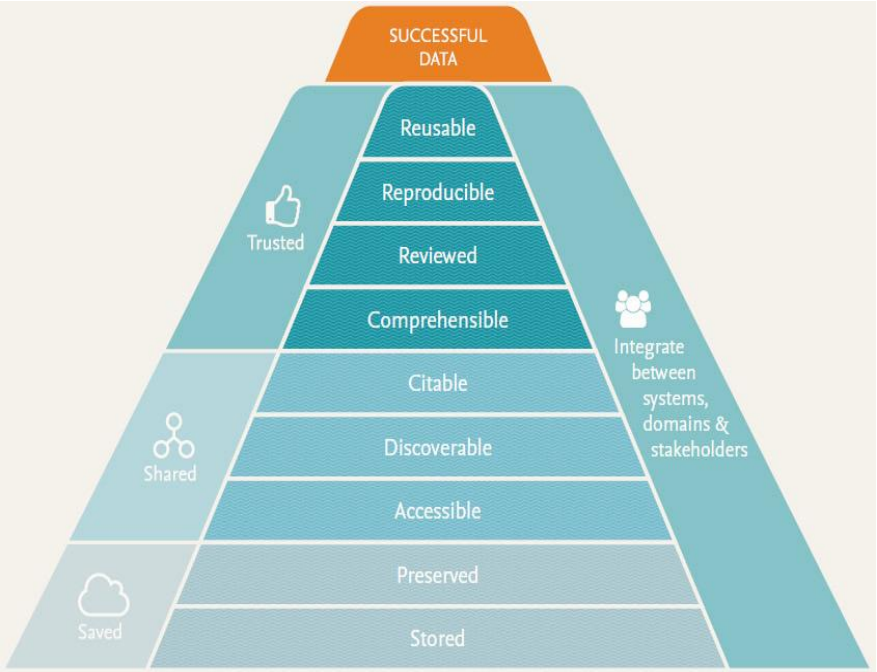
The goal is to move these objects from the supplementary storeroom and put them in the window, right next to the research article. There are at least nine steps to making optimum use of data and software, starting with storing and preserving it. The next step is to make it accessible,

discoverable, and citable. Once potential reusers locate the data, they'll need to have an idea how trustworthy it is. The criteria here are peer review, verification of reproducibility, and reusability. Ideally, information on all of these aspects would be available in the metadata for the object: Where is it? Where do I get it? Who gets credit? How did it review? Can I run it in an emulated environment, or do I have enough information to set it up myself?

Elsevier's approach has been to create what they call a "research element" (also referred to here as an artifact or an object): data, software, methods. Research elements are all citable, all peer-reviewed, and all published together with the primary paper. Research elements are mostly open-access, published under the Creative Commons CC-BY license, which lets users redistribute and build upon the work as long as they credit the original authors.

The idea is to reproduce the entire research cycle, to capture each stage in the life of an experiment, so to speak: planning; laying out materials and methods; writing simulation software; conducting the experiment; collecting, analyzing, and interpreting the data; and completing the experiment by publishing the paper and research elements. This experiment then becomes the foundation for more work, by the original author or others, and the cycle begins again.

Steps to Successful Data Curation



To get the maximum value from scientific data, it must meet nine criteria to be properly saved, shared, and trusted. (E. Presani, Elsevier)

A word about data, methods, and software. *Data in Brief* is dedicated, as the title suggests, to data. "Articles" are nontraditional, i.e., they consist of just the peer-reviewed data, the metadata needed to make it accessible, and a very short templated description to provide context to the reader. The idea behind *Data in Brief* is to make it very simple to submit data sets. The journal offers submission templates. And review templates reduce the bureaucratic overhead for reviewers. The data is made publicly available, again under a CC-BY license, with a description of the experimental design and references. When the data article is connected to a research paper, the link and citation appear in the article itself. For some formats of data, there may also be a small widget next to the article that will provide a simple visualization.

Elsevier's *MethodsX* captures protocols in a searchable, XML-based format. It embraces both methods that have been included in the materials and methods section of published papers and methods that may not have been cited in published research. It is dynamic, allowing members of a broader community to record modifications of the method (a change in temperature or initial conditions, for example) and note what impact these changes have on the results.

Software can be archived, searched, accessed, and cited via journals like *SoftwareX*. The journal publishes software connected to published research. The journal supports versioning, allowing contributors to freeze and record a version associated with a publication or a significant improvement. Each version of the software is stored in in *SoftwareX´*s GitHub repository, while the journal includes the bibliographic listing (making the code a citable primary object) and tracks citations, comments, software forks, downloads, and user ratings. Authors who already use GitHub can simply provide a link to the existing repository. Because GitHub users can delete their own repositories, *SoftwareX* caches a backup copy for these contributors so that the published version will always be available.

Finally, another journal, *HardwareX,* has also been launched, to capture and make available information about the hardware environments underlying published research, information that may not be captured in traditional articles.

None of these are ultimate solutions, and new systems will likely evolve for reviewing data and software. There may, for example, be indicators of trustworthiness, similar to ACM's reproducibility badging.

Group Reports on Peer Review and Quality Assurance

Nontraditional types of research products (e.g. experimental protocols, data, and software) will become increasingly significant components of the curated engineering research record, and the research will become increasingly "versioned."

New forms of peer validation; are persistent links possible? (Report from Group A)

The questions for discussion:

- *With virtually every product of research being updated on a continuing basis, what new forms of peer validation will be needed?*
- *Will it be possible to have persistent links between published papers and supporting software, experimental records, and data?*
- *Other — please identify any other items we should consider in the future.*

The group challenged its first question, "With virtually every product of research being updated on a continuing basis, what new forms of peer validation will be needed?" After 30 minutes of discussion, the group concluded that the question was "difficult to conceptualize in a world where we review fixed publications in fixed public versions." The group separated the question, to address validation of (a) software and then (b) other kinds of artifacts.

The open-source software community offers a good current example of continuous improvement and continuous validation. The group suggested experiments with crowdsourcing on validation of other kinds of artifacts, such as data sets.

Post-publication public annotation has proven controversial, or a failure, after 15 years of trying. Perhaps, though, open annotation might work if repositories invite open validation annotation via links between artifacts and the published peer-reviewed paper. These peer validations might not be blind, but rather credited, and might even involve collaboration between reviewers and authors.

The citation of an artifact on the publishing platform would change when there are updates to the artifact in the repository. Would this system wind up looking like CrossMark, and could CrossMark be the answer?

Because artifacts are continuously changing, a good approach would be to freeze and capture snapshots of a particular version, and have reviews address particular snapshots, rather than the main project in flux. Authors might question this, asking, "Why should I update the artifact when I can create a new one and get a new publication out of it?" Since new publications and citations are the currency of promotion, it is not clear whether communities would buy into programs of continuous improvement.

Should we advocate integrating publishing platforms and repositories? For example, the citation of an artifact on the publishing platform would change when there are updates to the artifact in the repository. Would this system wind up looking like CrossMark, and could CrossMark be the answer?

Question two was, "Will it be possible to have persistent links between published papers and the supporting software, experimental records, and data?"

"Yes," was the short answer. For published projects, many researchers are already working in the ecosystem of the RMap shared data site. As-yet-unpublished projects, though, might be registered at the institutional level. Each project could be assigned an ID. The institution would create metadata and identifiers and assign DOIs that would allow the researchers to begin creating artifacts, such as electronic lab notebooks for projects, as part of their research.

The question arose, "Where is the publisher on all of this?" Is this prepublication environment an opportunity for them? Or is the research institution where the work is done responsible for DOI assignment, management, registration, and so on? Today, the traditional article is an advertisement for the researcher's work and a promotion and tenure evaluation tool. If an internal prepublication system were to become available, would it, too, be used to grade the institution's researchers, once it's possible to track those projects at the institutional level? This type of information should not be exposed outside the institution: research should still become public through publication of a peer-reviewed paper and the associated artifacts.

During the questions, a participant from another work group noted that they, too, had talked about using CrossMark in new ways, saying that versioning and updating artifacts would likely appear as one of the meeting's common themes.

Another questioner asked what "continuous validation" means. Is there a timeframe for the validation? The answer was that there was no specific window of time. The artifact is exposed (published) and there is a mechanism for validating, almost like blogging about, the artifact around that particular time.

A third participant commented that the questions, "Is there a shelf life? Should there be planned obsolescence?" come up often in the contexts of reproducibility and verifiability. Does it make sense 20 or 30 years after publication to still be saying, "This artifact is validated and verified. It still works." After 10 years, say, an artifact could be demoted.

Software challenges; curation and quality engineering (Report from Group B)

The questions for discussion:

- *What are the greatest challenges we face in research software curation?*
- *What quality engineering practices are most critical for ensuring reproducible research software, and how do we ensure these are proactively implemented into our research?*
- *Other — please identify any other items we should consider in the future.*

The group amended the questions to include data challenges, to be able to tackle software and data together.

Greatest challenges. Supply-chain management is an issue. On one side, researchers are doing the research and producing research artifacts. On the other side are the consumers of the research. (Consumers might include other researchers, funders, industry, and others.) Aligning what's done on the production side with what's done on the consumption side would make for a smoother flow. Today, a researcher who wants to be able to publish artifacts as well as a traditional paper has to do additional work; eliminating that extra effort would be useful.

Curating software raises legal and commercial issues: licensing, intellectual property, copyrights, and so on. Even if researchers submit code they've developed on their own, and to which they hold

all the rights, the research may also rely on third-party software — programs, libraries, etc. — that they do not own, and which they are not free to redistribute under the licenses they hold. This creates a backlog of unsubmitted artifacts as these issues are worked out … or are not worked out.

Versioning is a challenge. What is the "official version" used in a research program? When should the code be released? If the research paper is submitted before the software is, the software may have continued to develop. Which version does the reproducibility reviewer review — the code used to produce the published results, or the best and latest version? The group discussed a possible two-stage process in which the research and software are reviewed as separate entities.

Scaling is a challenge. Peer review entails massive effort. Expanding the existing peer-review process to include software reviews would be very difficult. Reviewing software isn't like reviewing a paper. It's even more time-intensive. Reviewers are unlikely to be able to take other researchers' code, rebuild it and the environment on their own, and have it run correctly the first time. A single reconstruction is time-consuming. Repeated reconstructions are very time-consuming. Finding qualified reviewers who can actually do that job is going to be difficult. The pool of artifact reviewers will very likely be different from the pool of peer reviewers. This might be an opportunity. Or it might be a problem.

Software isn't always pretty. Getting researchers to submit code in the form in which it's used in the experiment is likely to be a challenge. Asking somebody to "show me your code" is the equivalent of saying "let me see you naked." That may be a hurdle.

Software isn't always pretty. Getting researchers to submit code in the form in which it's used in the experiment is likely to be a challenge. Asking somebody to 'show me your code' is the equivalent of saying 'let me see you naked.'

Quality Engineering Practices. The second question was, "What quality engineering practices are most critical for ensuring reproducible research software and data? How can we proactively implement those during the research phases?"

Peer review and code review headed the list, of course. Introducing quality engineering up front, early in research, makes it easier to prepare the package for any form of submission. Standardizing on a collection of platforms would help, along with the development of tools to help in the standardization. This ties in with configuration management: We need not just an understanding of the software version, but also an idea of what platforms were used, what tools were used, what libraries were used, what operating systems and languages, what build procedures.

This is a complicated process. Researchers need to be encouraged to submit artifacts, not discouraged by criticism during the review. There should be *no* shaming.

In a sense, the peer reviewers need to have some form of a "build package," allowing them to replicate the experiment on their own, or as independently as possible. The group used "containerization" to describe this packaging process, and some form of containerization tools or standards would be useful to have. As noted, better software development practices during research make the work required to submit software much easier.

Developing data standards is critical for promoting reproducibility. Later users don't really know what they're getting until they open a data file … and, too often, they don't really know even then. Even the simplest annotations can help: the units of measure used, the definitions for columns and rows, field-by-field dictionaries of codes used — "essentially a 'secret decoder'" to help the reviewer understand the file and the manipulations that produced it.

Other questions arose:

Should authors own the responsibility for appropriately curating software and data? Should publishers? Should the "reviewing entity"?

How should links between articles and software/data artifacts be created? Should they be submitted simultaneously or separately — considering that gathering the permissions needed to release artifacts could take considerable time to resolve licensing or legal issues?

Organizing and paying for quality assurance; models of peer review (Report from Group C)

The questions for discussion:

- *As scholarly research products move beyond the traditional paper to include data and code, the peer review process will also have to evolve.*
  - *How will quality assurance be organized? How will it be paid for?*
  - *How many reviewers are needed for high quality?*
- *What models of pre- and post-publication peer review will engage the community at a larger scale?*
- *Other — please identify any other items we should consider in the future.*

The group felt that the questions should be considered together. In particular, the artifact-review process (as outlined by Rous) raises questions of scalability. Will this labor-intensive approach work if applied more widely to software? And could it survive further expansion to cover reviews of data, methods, and other kinds of artifacts? How will it be possible to get enough reviewers when it is already a challenge to find peer reviewers for traditional papers?

The group attached a proposition, "More papers = more bad papers, going to the same pool of reviewers." The resources required — in people, time, and money — don't scale.

The group reviewed some successful existing models, evaluated in post- and pre-publication pairs:

- Funded mechanisms for post-publication peer review that allow a research community to cull the field and highlight areas that merit further research.
- Developing tools and infrastructure to start intramural repositories, in which artifact validation can be begun when the artifact is created.

The consensus was that there is room for both models. Again, one size will *not* fit all.

- ACM has had success with post-publication Artifact Evaluations, driven by graduate students and postdoctoral fellows motivated to learn new techniques and earn credit through the badging program. Artifact Evaluations are always done by at least two reviewers, and differ from ACM's Replicated Computational Results, which are formal review articles with full publication credit.
- Some disciplines (e.g., crystallography and cytometry) have established workflows for pre-publication data validation. Data goes into the discipline's repository, and the

researcher must take certain actions before taking that material through the peer-review process.

In general, researchers and students are investing time in trying to recreate algorithms for their own use, creating a pent-up demand for validations that will make the process easier. As a group, these early-career researchers have been very keen to get involved in the artifact-review process, attracted by the opportunity to get some credit, though there was concern that postdocs and grad students would only want to deal with "the sexy stuff."

Credit, Kudos, and Whuffie. One member of the group (Lynch) cautioned that this was a very publisher-centered approach. Much current research simply isn't interesting enough to recreate. And it is dangerous to build a process that is already overtaxed (in time, effort, and expense) into the publication process.

The group considered open or crowdsourced post-publication review, similar to Faculty of 1000 (F1000), reliant on volunteers. The incentive structure would be based on "kudos, credit and whuffie." (*Whuffie* "is the ephemeral, reputation-based currency of Cory Doctorow's science fiction novel *Down and Out in the Magic Kingdom*..." — Wikipedia)

Are there incentives to scale up an approach like this? Which is to say, "Can a researcher build a career on this type of validation?" The answer is, "No, not at a tier-one university on the tenure track."

And again, there's the question of who will pay for the expanded reviewing. The magical solution is to say that the funders need to support it (either from their very narrow operating margins or by reallocating funds ... in essence, reducing the amount of research to increase the quality of the research that is done). But is there a market-based approach that would allow scientists and engineers to do quality assurance rather than more peer-reviewed research, which is what drives the rest of the academic reputational economy? Could they get kudos but also some kind of financial benefit — extra funding or discounts on publishers' article processing charges (APCs)? It might be possible for researchers to sell "bonds" on their research, futures-market-type bets on the quality of their research. Is this worth talking about?

Funding needs to consider the totality of the research process. Should there be a pool of cash reserved for reproducibility preparation and review of particularly good work?

In the end, funding needs to consider the totality of the research process. Should there be a pool of cash reserved for reproducibility preparation and review of particularly good work? Researchers could apply separately for funds to bring artifacts up to spec, and third parties could then do the work. It's still the same pool of cash at the end of the day.

There is an assumption that researchers will not want to participate in reproducibility validation, if the efforts must be paid for out of existing funds.

A participant likened the dilemma to an old Irish joke. "How do I get to Dublin?" "Oh, I wouldn't start from here." Much of the thinking in this area starts from where we are now, and it may be a real impediment to thinking about where we need to go.

Different environments vs. a common platform; addressing proprietary code (Report from Group D)

The questions for discussion:

- *As nontraditional types of research products (i.e., data and software) become a significant component of the curated research record, how can we run experiments using different software and environments? Do we need to provide a common platform?*
- *Do we need to address the possibility of proprietary software (e.g., compilers) and environments? If so, how do we address proprietary software and environments?*
- *Other — please identify any other items we should consider in the future.*

Should different software and different data all go on one platform? Do we need to provide a common platform? Developing a new initiative is a three-step journey, from adoption to acceptance to perfection. Developing a common platform belongs more to the "perfection" phase. It may be premature to take it up now. The question also brings up the difference between replication and reproducibility.

For software and data, replication means that you want to keep artifacts together. To confirm reproducibility, broadly defined, users and reviewers will want to be able to test the artifacts separately. One researcher might say, "Let me test that software on my data," and another might say, "Let me test my software on that data." This is where reproducibility questions become far more interesting: one can see how robust certain artifacts are, and whether they remain reproducible under small changes in inputs or environment.

To what extent does proprietary software spoil the game of sharing and openness? There is a general notion that the results of publicly funded research should be open and not proprietary. But there are also important proprietary goals for stimulating technology transfer and catalyzing new businesses.

In that sense, it will probably be for the good of science *not* to have one platform/environment, but many, even "to let a thousand flowers bloom." There is a counterargument that combining many diverse data sets on a single platform lets the researcher look for far more interesting patterns and relationships, transforming "a lot of data" into "big data." On the other hand, there's, of course, also the issue that if you can't combine a lot of data sets, that you can look for far more interesting patterns and relationships and, in that sense, make more data into big data. There are, of course, complexities that must be solved before you can do that. Again, the oft-used word "standards" comes to mind, especially interoperability standards.

Interoperability standards would also apply to platforms. In our current phase of development, it's more important to set standards for operability, rather than trying to build or specify a platform. Researchers need freedom, and a one-platform environment can be very restrictive.

The second question concerned proprietary software. To what extent does proprietary software spoil the game of sharing and openness? There is, of course, the general notion that the results of publicly funded research should be open and not proprietary. On the other hand, there are also

important proprietary goals for stimulating technology transfer and catalyzing new businesses. There is certainly a tension between these two goals, which is important to the funding discussion.

Assuming that proprietary software is in use, and will continue to be for some time, it is important to develop software-test techniques that are fairly simple, and can operate without having to be able to read *all* of the code. Such techniques would likely be faster to develop and faster to run. Being able to develop a measure of trust in the software is very important.

The catchall "Other" question sparked a lively discussion about the peer review of data and code, and how the mechanics might overburden and jeopardize the whole peer-review system. Artifact reviews might take too long. The usual peer reviewers might not be the best positioned to undertake artifact review. *Et cetera, et cetera*. There were ideas about how to address the problems — not immediate solutions, but worth thinking more about.

There is a movement afoot to build — and NSF has spent a lot of money to develop — the 'data commons,' in which a community of researchers is responsible for creating a common set of data that can be widely used through the broader community.

Several suggestions came up in the discussion. They were not immediate solutions, but bear further consideration. There was a suggestion to take the approach DARPA takes, where the process is not so much a peer review as a collaboration by what they call "challenge partners" — peers who examine the software and bring back suggestions for improving it. This does take time and effort, and raises the question, "How does that get funded?"

The group noted zero-level testing, a quick run of the same software on the same data, to confirm that it produces the same result. There was also a crowdsourcing suggestion emphasizing post-publication peer review, "just push the artifacts out and see how the community reacts." Once again, the discussion returned to trust and versioning. CrossMark was again mentioned as a tool.

Baillieul underscored the point that there is a tension in making shared data and software accessible to the community outside of the traditional publication processes. Researchers *do* like their freedom, and they *are* going to choose their own code and gather and archive their own data sets. On the other hand, there is a movement afoot to build — and NSF has spent a lot of money to develop — the "data commons," in which a community of researchers is responsible for creating a common set of data that can be widely used through the broader community. There is a big challenge in thinking out the trade-off between uniformity and giving researchers enough space to develop as they see fit.

Plenary Panel: The Economics of Reproducibility

As the current scholarly publishing business model undergoes pressure from the tilt toward open access, and library budgets are further reduced, how will the added step of reproducibility be funded? Panel will discuss funding scenarios.

Panel moderator: Gianluca Setti, Department of Engineering, University of Ferrara, Italy.

Panelists: Todd Toler, John Wiley & Sons; Jack Ammerman, Boston University; Victoria Stodden, University of Illinois – Urbana-Champaign; Dan Valen, Figshare.

Digital First: A Publisher's Value Proposition in the Age of Reproducible Science and Open Data

Todd Toler, vice president of digital product management at John Wiley & Sons.

It is hard to think about the economics of publishing without thinking about what's coming next. Consider, as a basic thesis, the proposition that future initiatives in reproducible research would be author-funded through data processing charges (DPCs).

From the funders' point of view, "there's just not that much more money in the system to … go toward a whole second suite of [data processing] charges." As Toler sees it, there are two "flavors" of DPCs: an extra charge levied on top of others, like the print-era color and page charges, or a charge for double publication, in the core journal and a data journal.

In the first model, assistance in data curation and validation is a value-added service, covered by the data processing charge. This follows the model of the page and color charges, revenue streams to which many publishers are addicted. Note, though, that these charges are increasingly unpopular and they make less and less sense as the focus on print weakens.

The review process is the soft underbelly of ambitious plans for near-universal data and software validation. There are not enough postdocs and grad students in the world to take on all of the work that needs to be done.

In the second model, the DPC is a "double publication charge." Yes, the authors pay a second article processing charge, but they also gain a second, linked, citable publication. Data journals are on the rise because existing journals lack systems for data attribution, or micro-attribution, or data citation. Instead, one sees data-providers receiving co-author credit rather than being cited directly. The co-authorship is their incentive for sharing.

Data-only journals may be an imprecise, stop-gap response to the challenge of data storage and curation. The data journals have the focus on data, but it's not going to scale. There is ample room and time for journals to move into the field, developing the "omics" approach and simulation capabilities of data-intense sciences generally.

The review process is the soft underbelly of ambitious plans for near-universal data and software validation. Toler pointed out that there are not enough postdocs and grad students in the world to take on all of the work that needs to be done.

Instead, the transformation will have a couple of elements. Most important is a switch to a publishing paradigm based on data, not on the printed word.

There needs to be more cooperation and agreement on standards — file formats, interfaces, processes. At a recent Library Conference discussion about SciHub, librarians pointed out that SciHub exists because publishers have failed to solve some long-standing problems. Publishers have convinced themselves that they are competing on platforms, and that they have to differentiate themselves. That isn't what their library customers are asking for. Standardization presents vast opportunities for efficiencies across the entire publishing infrastructure, from initial submission and review through to the user experience.

Funders, overstretched as they may be, do have a role in this process. NSF, for example, is making some very targeted grants in the cyberinfrastructure area. Strategic investments in the data-and-software infrastructure will really help move the field along.

Finally, though, the onus will be on the publishers, Toler said. "We're living in a GitHub world." GitHub claims more than 19 million users (as of January 2017), and scientists are its fastest-growing segment. Toler says that Wiley has "a whole floor of Ph.D. editors, and every one of them can write Python code."

The point is that many of today's researchers have been raised in a world of open outputs, online collaboration, version control, and software that can be forked, branched, and merged. "The idea that your data will be sitting on your hard drive and not linked to your research output just doesn't make sense in a world where you grew up basically collaborating on the internet."

The page is still the basic unit of content. Publishers still do their budgets in pages. When they work with vendors, prices are based on pages. Page charges are levied on authors for long articles, as though to cover the costs of extra printing, paper, and postage.

Publishers need to rethink their workflows, rebuilding them from the ground up for the web. Scientific journals started going digital in the early 1990s; the widespread adoption of the World Wide Web and browser technology around the middle of the decade led to a very rapid shift to acquiring and using journals in digital form. By the late 2000s, university libraries were increasingly discontinuing their print subscriptions (after a few years of transition during which they acquired both print and digital).

Today, 95% of the value of a journal comes from digital distribution, site licenses and the like. Print remains a strong factor only for certain society journals and journals that generate significant advertising revenue, for the most part health and biomedical publications

Yet publishing remains focused on print-like objects. "We still basically have a system that's a digital lens over a print-based system." Even though almost all distribution is digital, the workflow is still structured as though they were print product. The page is still the basic unit of content. Publishers still do their budgets in pages. When they work with vendors, prices are based on pages. Page charges are levied on authors for long articles, as though to cover the costs of extra printing, paper, and postage; additional editorial or infrastructure costs caused by complexity are not yet part of the equation.

After the review process (itself still sometimes heavily dependent on PDFs, faithful images of typescript), the process focuses on typesetting the article, which is expensive and slow. The

searchable electronic version is created almost as an afterthought, from the print image. "There is a huge opportunity to just create a better workflow for authors, [one] that will just organically lead to a more reproducible scientific publishing infrastructure."

Here is the process as Toler outlined it. The author submits a Word document, image files, and other material through an online submission application that "looks like it was designed in the '80s." For the next few months, the package shuttles around the peer review process as a collection of files and attachments. At this point, the files are sent to an offshore article-making factory.[c] There, the articles are composited, typeset, into an XML file (generally NISO JATS, the National Information Standards Organization Journal Article Tag Suite[14]). The XML is printed out into a PDF, which is sent back to the author for revision. The author then inserts corrections as comments on the PDF. The corrections go back to the typesetter, who (manually) transcribes the changes into the XML file. Then another PDF, of the final pages, goes back to the author for sign-off (and, too often, last-minute changes, which the typesetter must again transcribe). The typesetter then sends the article, as a PDF and XML to the publisher. And the publisher loads the PDF and HTML (HyperText Markup Language) derived from the XML into the content database. This is the workflow, *even if the journal never goes to print at all.* At an average typesetting cost of $6 to $7 per page, and a turnaround time that averages 22 days, the process is both expensive and slow. Worse, typesetters can't handle data, and, ultimately HTML is semantically weak and not very useful on the internet.

In this case, most of NISO JSAT's semantics concern content layout, with a smattering of scientific relevance.

"This XML is not an open web standard," Toler said. "Google does not speak NISO JATS. Google speaks JSON LD (JavaScript Object Notation for Linked Data[15]) and Schema.org. Google speaks open web standards that are worked on by the W3C. All search engines do."

With the rise of the semantic web, with schema.org and search engines agreeing on what should constitute machine-readable metadata under the human layer of HTML, something new started to emerge. *The New York Times*, for example, tried including JSON LD metadata web snippets with cupcake recipes. Their web discoverability went up 50% on cupcakes.

So, today, cupcakes benefit more from modern data practices than the entire scientific enterprise, which is still wedded to HTML produced by typesetters.

"This is not good for reproducible science," Toler said. "I picture a world where we have scholarly HTML. Under that, there's a JSON LD metadata layer [giving scientific context, and] behind every figure is a link to the repository where the data is sitting."

Publishers need to figure out a way to publish and link data without destroying its value by typesetting it. Print, or print-on-demand, can come later. The journal should be able to take a submission as HTML or XML or iPython or Jupiter Notebook and knit elements together organically. The "research paper" part might include discussions and conclusions and references,

[c] According to Toler, offshoring is not the issue. The issue is that publishers have elected to drive costs out of the print process rather than seek new processes that are inherently more efficient.

and publishers would add quality control and integration, but data and software would not pass through links so much as "just work," through direct pipelines to the original stores and tools.

Wiley is building such a system. PLOS is building another, Aperta. Other publishers are also working on online-first HTML processes. So authors will submit their Word manuscripts, and figures, and tables, and other material. The submission will be converted immediately into HTML, and editors, reviewers, designers, and the authors will then all work on the same HTML document.

The referee would no longer review a Word document and have to connect a figure reference in the text with a link to a JPEG in a data center or a bit of code. Instead, the reviewer would do as any end reader might, and click on the figure in the article to immediately display and/or execute the code behind it. The referees' and editors' comments will be integral annotations and part of the web content.

Hypothesis (hypothesis.is) is an open-web annotation service doing just this kind of thing. Ultimately, reviews will be a part of the web article package. Based on the individual journal's policies, the annotation server will determine whether or not the general reader can see it. This is what Hypothesis is doing.

In this integrated submission system, the author becomes a partner in raising the quality of the submission. After the publisher converts the article to HTML, the editors, reviewers, and the authors can start defining the quality control approach. Does the author want to make the work more reproducible? Is the submission complete? Is it semantically well described? Is there data or code behind the figures? The result is a reproducibility and quality score that rates the presentation's completeness and quality. It does not judge the science.

The system puts a heavy burden on researchers and reviewers. Automating quality and completeness tests (not the evaluation of the science, but of the formal suitability for web publication) can help ease the load.

The system puts a heavy burden on researchers and reviewers. Automating these quality and completeness tests (not the evaluation of the science, but of the formal suitability for web publication) can help ease the load. The first reactions are likely to be negative ("Robot referees? Really?") but the productivity improvements will win acceptance, by eliminating much of the "editorial drudgery" tax on the whole research enterprise.

Breaking the "Iron Triangle": Perspectives on Sustainability of Institutional Repositories

Jack Ammerman, Associate University Librarian at the Boston University Libraries, has long experience with library technology, policy, and governance.

The library places a central role in the information ecosystem. Research librarians have been collecting "research output" since the invention of libraries. Until the mid-1990s, these outputs were books and journals. When online publications emerged, librarians added them to the mix of objects to be collected, preserved, curated, and described in metadata that made them discoverable. Librarians are only just beginning to think about larger, more complex research

objects. Libraries are digital and networked, with a global notion of what their services and collections should look like.

A library budget is allocated among three primary areas, in an "iron triangle": content, infrastructure, and services. If the library adds another subscription, that has an effect on infrastructure and services (e.g., increasing demand while reducing the pool of available money). If the library adds librarians to provide additional services, this affects the funds available to buy content. The steady, large increases in the cost of commercially published journals — the journal crisis — has affected libraries' abilities to buy books; it has also eroded the other services the library provides.

Added to discussions of limits on research funding, research capacity, and reviewer capacity, this iron triangle seems like a symptom of a model that has librarians, researchers, funders and publishers locked into a zero-sum game, particularly with budgets that have stayed flat for five or 10 years. It's an information ecosystem in which all of the stakeholders depend on one another for survival, even as they fight for larger shares of a fixed pie. So they have come to view one another as adversaries as much as partners. All of us, it seems, are trying to find a viable business model.

If stakeholders cling to their traditional working relationships and practices, however, incremental reforms and efficiencies will still leave them bound in the same iron triangle. Breaking out will require rethinking the mission, eliminating some practices and adopting some new ones. Boston University's libraries have been streamlining workflow, dropping print subscriptions, and rethinking traditional mandates and management practices built around print holdings. Increasingly, the libraries are integrating with external services to reduce the kinds of duplication seen in the past.

Boston University's libraries have been streamlining workflow, dropping print subscriptions, and rethinking traditional mandates and management practices built around print holdings. Increasingly, the libraries are integrating with external services to reduce the kinds of duplication seen in the past.

BU is reviewing services, some of them traditionally considered indispensable. The libraries will likely eliminate its circulation desk in the near future, and replace it with a self-checkout system lightly staffed with "student ambassadors." They are modifying the focus of acquisitions; rather than trying to forecast what users will need for the coming year or decade, the library will explore a patron-driven system, purchasing products when they are requested by users. The result, said Ammerman, will be a broader array of materials available to the user … without spending all of the acquisition budget up front. The library is redeploying staff: As the circulation desk is wound down, it is hiring a new data services librarian, adding support for digital services, and developing collaborative relationships with other libraries and information sources.

Lorcan Dempsey, vice president for membership and research and chief strategist for OCLC, is a proponent of "right-scaling." Libraries have a history of treating their problems as though they are peculiar to their own institution. They invent local solutions. Instead, Dempsey suggests, librarians (and by extension, others involved in the research enterprise) should think more deeply

about the issues they face, understand the scale of the problem and the extent to which others share it, and then address it at the right level: a local group of libraries, a regional consortium, or a global initiative. Solutions should be devised at the appropriate scale.

An October 2016 Massachusetts Institute of Technology Task Force on the Future of Libraries report refers to the "library as a platform."[16] The concept is consistent with what Ammerman sees as Boston University's direction: building APIs for the library; making the library "hackable" (in the best sense); publishing bibliographic records on the web in linked data formats, making the BU holdings discoverable far beyond the school itself.

Ammerman stopped short of suggesting a corresponding "researchers as a platform" approach, but he noted that there might be a kind of research platform that would embrace resources like Figshare or the Center for Open Science and others, integrating them, obscuring boundaries, breaking the iron triangle to provide the kind of services the information ecosystem needs to do reproducible science.

The Economics of Reproducibility

Victoria Stodden, statistician and advocate of reproducibility in research at the School of Information Sciences at the University of Illinois at Urbana-Champaign. The work she presented was done in conjunction with David Donoho.

There appear to be tandem ways forward, encouraging and expanding ability to do reproducible research, and expanding access to computation on every scale, including massive scale. The responsibility for making reproducibility a reality does not lie with any one group. It's a challenge of collective action that reaches down to the incentives of the individual researcher. Funding agencies, in particular, have a leverage that can be used to push the reproducibility movement forward. Stodden's thesis is that, through grant set-asides, these dual movements can advance hand in hand with what is being called a "reproducibility industry."

Transparency, verifiable reproducibility, and sharing data will allow researchers to run more experiments and larger-scale experiments more efficiently.

On the surface, some of today's important trends may seem antagonistic, as scientific projects are becoming more and more computing-intensive, while science is also moving toward transparency and reproducibility. These demand time and care, and some may see them as time-sinks, as a drag on research.

These trends may not conflict at all. In fact, though, transparency, verifiable reproducibility, and sharing data will allow researchers to run more experiments and larger-scale experiments more efficiently. And the computational infrastructure being built for ever-larger-scale projects will also promote transparency. So factors that seem antagonistic are actually mutually reinforcing.

Consider the National Institutes of Health's decision to require that clinical trials include biostatistics Ph.Ds to enforce rigor in experimental design and analysis. NIH funded the biostatistics capability through grant set-asides, transforming the clinical trials process, creating

a cottage industry of biostatistics, and, not incidentally, bringing the public more bang for its tax buck by delivering better science.

To support reproducibility, each grant could contain some amount (say $500) for each publication coming out of the funded research. This money is earmarked to pay for reproducibility certification by some accredited third party and an accessible deposit of the objects and information needed to reproduce the paper's results. Such a grant, then, helps fix responsibility for preserving and curating the elements of reproducibility.

The certifier might be a journal, a scientific society, a library, or some new kind of entity. The certifier would develop, or underwrite, reproducibility-related tools that authors could use in their research. These tools (like Code Ocean) have the dual virtue of promoting both reproducibility and efficiency. The consensus is that there is no one-size-fits-all reproducibility solution. The proposed distributed approach would provide incentives and a mechanism for adapting certification standards to the requirements of each discipline and subdiscipline.

Standards and tools will naturally evolve over time. Much of the reproducibility certification process might be automated.

Reproducible research should also be more fully annotated, allowing much more sophisticated searches that are possible today. Future systems might be able to answer queries that are unanswerable today, such as:

- Show a table of effect sizes and p-values in all phase-3 clinical trials for melanoma published after 1994;
- Name all of the image de-noising algorithms ever used to remove white noise from the famous "Barbara" image, with citations;
- List all of the classifiers applied to the famous acute lymphoblastic leukemia data set, along with their type-1 and type-2 error rates.

Starting in September 2016, JASA ACS (the *Journal of the American Statistical Association: Applications and Case Studies*) added to its masthead three reproducibility editors (including Stodden). At the same time, JASA ACS began to "require code and data as a minimum standard for reproducibility of statistical scientific research." Commitments to provide code and data post-publication are often fragile. The JASA ACS requirement and its reproducibility editors help ensure that these materials are gathered and tested before publication. As a rule, statisticians use standard statistical languages, so the editors expect the problem to be tractable, with few challenges from unusual hardware, odd computing environments, and the like. In conclusion, Stodden said:

> I think it's clear that we should proceed assuming that we are not going to get additional funds to do this and that we should work within our initial set of current financial constraints. However, I think we need to bear in mind that the reason we are doing this is for notions of transparency, for access and to maximize the ability of the community to verify the work. So I mentioned that because through the discussion yesterday, we saw people saying IP is very important, access is very important.

The Road to Reproducibility: A Look into Research Data Management Initiatives and Researcher Behavior

Dan Valen is a product specialist for Figshare, "a repository where users can make all of their research outputs available in a citable, shareable, and discoverable manner."[17]

Governments in North America, Europe, and Australia have unanimously agreed that the trend toward reproducibility and openness in research is a good thing, and something they need to pursue.

In October 2016, the Center for Open Data Enterprise published its *Open Data Transition Report: An Action Plan for the Next Administration*.[18] The white paper describes how to continue the momentum toward use, reuse, and republication of open government data … including efforts to "empower researchers to advance scientific discovery and drive innovation."

This is just one of many working groups, interest groups, and initiatives around the world promoting transparency in information.

Another is the Research Data Alliance (RDA), a non-governmental organization launched in 2013 with support from the European Commission, the U.S. National Science Foundation and National Institute of Standards and Technology, and the Australian Department of Innovation "to build social and technical infrastructure to enable open sharing of data."[19] RDA has interest groups and working groups, and any interested party can join. They are thinking about many of the same issues addressed at this workshop, Valen said. And RDA is aggressively focused on the economic benefits of open research. Consider two examples: A study of U.S. spending on the Human Genome Project estimated that the $13 billion the government invested returned an economic benefit of about $1 trillion. And in the UK, every £1 invested in research returns economic benefits of about £5.40.

A study of U.S. spending on the Human Genome Project estimated that the $13 billion the government invested returned an economic benefit of about $1 trillion.

CODATA, the Committee on Data for Science and Technology of the International Council for Science (ICSU) was founded 40 years ago to work to improve the quality, reliability, management and accessibility of data in all fields of science and technology.[20] Among the issues CODATA currently addresses are the sustainability of data-repository business models — commercial, like Figshare, or grant-funded, library-supported, or open-source community-supported — and the crucial question, "Which revenue strategy will ensure that the content will remain open and available into the future?"

The Illinois Data Bank (IDB), "the public access repository for publishing research data from the University of Illinois at Urbana-Champaign," represents another approach.[21] IDB was built from scratch and launched in 2016 after a year of development. The project is funded not by the UIUC library, but by the university's vice provost's office. The research office is paying close attention, and the libraries are providing strong support — expert advice on research-data management and the services needed to build good metadata and ensure that the content is discoverable and durable. In a recent publication, IDB's developers wrote:

> …if history were to repeat itself, in five or 10 years we can anticipate that agencies will develop their own data repositories with associated requirements that researchers deposit their federally funded data within those specific agency-led resources; thus, we have prepared ourselves for the chance that our efforts to build a data repository may be a short-term, stop-gap solution.[22]

Valen noted parenthetically that this contradicts some forecasts made earlier in the workshop.

A report commissioned by the Wellcome Trust attempts to hash out the costs and labor required to clean, prepare, and format research metadata. What does that necessarily involve? What problems will these overheads pose for libraries, researchers, or publishers? [23]

Valens noted that *Science* magazine, the Open Science Framework, and the *Public Library of Science* (PLOS) have all issued transparency guidelines.[24,25,26] In 2014, when PLOS announced it would start to require open access to data associated with published papers, researchers resisted. Today, he noted with satisfaction, open access is no longer frightening, and a diverse group of stakeholders in scholarly publishing are discussing how to support reproducibility and access to data, rather than resist it.

Researchers can have a potent disincentive to sharing hard-won data. A scientist can mine a good data set for a number of publications, earning citations that count heavily toward tenure and promotion. While it clearly helps science to let the information out, other people's papers don't necessarily help the original authors keep their current jobs. Valen pointed to a possible new dynamic in the experience of Stephen Roberts, an environmental scientist at the University of Washington.[27] A 2013 profile of Roberts in the *Chronicle of Higher Education* described how he used downloads of his data sets (from Figshare) and other evidence of online impact as part of his tenure package. He won tenure. More recently, in a *Science* opinion piece, "The Hard Road to Reproducibility," Lorena A. Barba described the discipline her lab exercises in working toward reproducibility. In the article, Barba says,

> Every result, including those from failed experience, is documented every step of the way. We want to anticipate what other researchers might need to either reproduce our results with code or our data, or replicate them. [28]

Whenever the lab submits a publication, they also deposit their data — to GitHub, Figshare, Zenodo, the Open Science framework, or elsewhere — with pointers that lead back to the article.

Another 2016 article, "A Healthy Research Ecosystem: Diversity by Design" in *The Winnower* ("Founded on the principle that all ideas should be openly discussed, debated, and archived") envisions the larger ecosystem, with a multitude of players, but in which all of the objects and tools and infrastructure talk to one another. How can that be facilitated? Open APIs are the first essential. Users must be able to query all of the content, and do it programmatically. Open access is the second essential, facilitated by open licensing. This will ensure that the content will be made openly available and will continue to be openly available; no one can slam the door on the content.[29]

And a final citation: *The State of Open Data Report,* published by Digital Science and Figshare, a 52-page "selection of analyses and articles about open data."[30] Valen described it as "wide look at the current ecosystem and where we think it's moving."

Group Reports on the Economics of Reproducibility

Publishing professionals are exploring approaches to data and software curation in engineering and other disciplines. Data professionals who currently provide platforms for such curation as well as those engaged in research on fundamental data science, data infrastructure, and cyber-infrastructure are investigating ways to harvest value from curated research products. With the goal of ensuring that future engineering research will be maximally reproducible, how do we develop new advances in data infrastructure and analytics, reproducibility, privacy protection, and research in the human data-interface?

Stakeholder roles; at what scale do we address challenges? (Report from Group A)

The questions for discussion:

- *For each of the following stakeholders, what role do they have in greater research curation and what primary needs or concerns do they have?*
    - *University and institutional libraries*
    - *Journals*
    - *Funding agencies and government*
    - *Government leaders, such as OSTP and U.S. Congress, etc.*
- *At what level (local, group, national/international) does it make sense to try to address each need/concern?*
- *Other — please identify any other items we should consider in the future.*

These questions were the most hotly contested of the workshop. There were, however, some areas of agreement. It is hard to do justice to the first question, "For each of the following stakeholders, what role do they have in greater research curation and what primary needs or concerns do they have? University and institutional libraries; journals; funding agencies and government; government leaders such as OSTP and U.S. Congress, etc."

The group tried to map the flow of money, inconclusively. They did agree that many research services have grown out of universities and university libraries. It would be efficient to take further advantage of librarians' experience with best practices in data management and curation. This builds on Ammerman's characterization of the "library as platform." Growth might be managed through a large, open-source community, or through regional consortium agreements among universities banding together to provide support for their research communities.

Publishers might expand existing infrastructure and band together to establish standards. Some collaborations are already underway, both within technical societies and in commercial publishing. These include such organizations as the National Information Standards Organization (NISO), CrossRef, ORCID (https://orcid.org/), and CREDIT. Publishers might collaborate to support best practices for research data management — and thus for reproducibility, reuse, and replication of research data.

Disagreements began when the discussion turned to what role funding agencies and government thought leaders should play. The group agreed on the need for a system in which everyone is part of the process. There was a lot of back-and-forth about how to achieve this.

Some maintained that the funding agencies are the force that drives change, which will not, in any case, happen overnight. In the meantime, who takes responsibility for curation? The idea of

putting money behind reproducibility and curation through grant set-asides had proponents and opponents. The discussion blended into the second question: on what level does it make sense to address these needs and concerns — local, group, national, or international?

Reproducibility is an international issue, and there are groups that do address it on the international level. Some countries have started to provide nationwide infrastructure (Australia and Japan, for example). The group discussed how these national efforts might be accomplished in the U.K., European Union, North America, Asia, and the Southern Hemisphere.

Does policy-making start from the top down, progressively narrowing to discipline-specific options or solutions? There was no clear consensus to be achieved in the time allotted.

There is the Research Data Alliance, which meets every six months. RDA working groups and interest groups focus on metadata standards, best practices for repositories, repository sustainability, and they do get down to discipline-specific concerns, such as geology and polar climate data, and best practices for managing that content.

Publishers should be involved in that process. RDA is open to scientists, non-profit organizations, and to the funders. It's a forum in which everyone can be part of the conversation, take information home, and figure out what value they can bring to the enterprise.

Eefke Smit commented on the activities of the Research Data Alliance and its working group on data publishing. RDA's definition of "data" embraces any kind of research output, including multimedia, computer code, protocols, methods, and so on. The large data working group has several subgroups. There is a subgroup on workflows, especially manuscript submission workflows in which authors also submit data sets. There is another subgroup on metrics for data citation. A third works on linking publications and data sets; RDA has just launched Scholix, a linking framework for data sets and publications. And a fourth subgroup addresses business models for data publishing. RDA is open to all who are interested.

What policy efforts are underway? How should we think about funding and resources? (Report from Group B)

The questions for discussion:

- *What ongoing policy efforts, in particular by different funding agencies, are underway to provide public access to engineering research products?*
- *How should we think about funding or getting resources to address curated research and reproducibility of research?*
- *Other — please identify any other items we should consider in the future.*

The group amended the first question, dropping the restriction to engineering to include "provide public access to research projects." The group agreed that the question would be best answered by research, rather than discussion: someone should find out what different agencies do.

The group nonetheless addressed the question. The main concern is that there are policy guidelines, but no funding. It is easy to put stipulations in place, but implementations can be difficult. It can also be bureaucratic.

The academics in the group were familiar with policies within NSF, but not within NIH. NSF is taking steps in the right direction: it has a data-sharing policy, but there was a concern that the

requirements didn't really have teeth. Grant recipients have to write and submit data management plans, but they are not always implemented. It's not clear what material should be saved. It's not clear how it should be saved. There is no quality control, and the process is generally decentralized.

It's a step in the right direction, but a short step. There were strong concerns that reproducibility requirements would become an "unfunded mandate." This drove the group to the second question, "How do you pay for curating research and validating reproducibility?"

Most of the ensuing discussion was about *who* should pay, which was easier to address than *how*. The group identified five models of who would pay — all of which came with caveats.

One, the government pays. The concern, of course, is that reproducibility and curation funds would come out of the same pot of money as research funding. There would be no overall increase in funds, so researchers would logically complain that funding data archiving would cut back on the amount of research that was funded.

Two, industry pays. Industry benefits from research results, but they don't contribute. Is there some model that we have where industry pays part of the cost? Take PubMed, for example: perhaps the pharmaceutical industry could chip in to support PubMed's infrastructure.

Or consider the Semiconductor Research Corporation, an industrial consortium that funds research projects, including some projects funded jointly with the NSF. Perhaps some of those funds could support data storage and reproducibility. What would industry's value proposition be, though? Industry groups might not be interested in collaborating, and if they are, they might be interested in only supporting work with an immediate impact. Still, this might be a place to start.

The group added a third question: 'Are there ways of harvesting value from curated research projects?' via a service that would aggregate artifacts, information, metadata, etc. Again, there is the question of who pays.

Three, researchers pay. This might be done through crowdfunding. Article postings could include a mechanism for requesting the research data. If the number of requests grows large enough, it would open a crowdfunding platform through which the researchers who want the data contribute to the data-deposit costs. When the promised contributions reach some threshold, the original researcher makes the data available. But where would the data reside? On a thumb drive? And if so, does that actually qualify as "preservation"?

Fourth, an independent foundation pays. Would a Carnegie Foundation or a Sloan Foundation underwrite big infrastructure projects? And is there a long-term sustainable future for projects begun this way?

Fifth, the publishers pay. The question here is whether publishers would simply pass the costs on to researchers, either as a publication charge or a subscription product. Publishing is a business, and the costs would have to be covered in some way.

The group then added a third question: "Are there ways of harvesting value from curated research projects?" via a service that would aggregate artifacts, information, metadata, etc. Again, there is

the question of who pays. There might (and this is just an example) be a Google or Google Scholar search option that limited results to quality-assured research. But then someone must decide what "quality-assured" means, and decide which research meets the criteria, and so on.

One participant noted that the "government pays" option was distinct from the suggestion, made by Stodden and others, that a certain percentage of grants be set aside for preservation, curation, reproducibility, and related activities. The idea was more that programs would target a specific area, or subdiscipline. They would create tools and infrastructure to support research, so that software and data are, in a sense, "born" ready for curation. This would be a specific project, a smaller-scale infrastructure moon shot. The project would be funded as a whole; it would not be a portion out of each research grant. It does all come from the non-expanding pool of government funding, but there is the potential for leveraging infrastructure in the way that one can't leverage individual vetting of a particular research object. This might pay dividends in the long run.

A participant suggested that investors might support reproducible research targeted toward that architecture. For example, Intel sells chips and builds resources so that they can sell those chips. If you can convince them that the computer-vision algorithms or machine-learning systems you're building are targeted on Intel architectures, they might buy it as a platform that would benefit their products. Companies often want to own that kind of research, rather than share it.

Another example was hardware manufacturers who give away software that enhances the value of their products. For example, Intel (a partner on several projects) doesn't care if an academic collaborator gives away the libraries they develop for Intel. The company cares about supporting its chips. The software becomes value added, an additional reason to specify their products for cars or mobile devices, or whatever. Users can do development better on Intel architecture. They like software that's reproducible — but on their architecture. A car manufacturer will now specify Intel architectures because these libraries can be reproduced on them.

Can we publish papers that solely cover research reproduction? Will non-RR papers cause legal issues for the publisher? (Report from Group C)

The questions for discussion:

- *Can papers that only reproduce results be published? Will publishing reproduced work lead to people who only do this?*
- *How can a lack of reproducibility be prevented from causing legal issues for the publisher? What legal implications and intellectual property issues might arise from artifact sharing (data, code, workflows, etc.)?*
- *Other — please identify any other items we should consider in the future.*

To answer the first part of the first question, the group agreed, "Yes," papers that only reproduce results will be published. First, there is a wide range of quality control and decision making in publishing. If these papers are created, they will probably find a home. On the other hand, a paper that confines itself to reproducing results may not be very intellectually compelling. If that is all that a researcher does, it probably won't help his or her career, and it won't earn tenure or promotion at a major research institution.

To answer the second part, the group did not see research devoted entirely to reproducing others' results as a [major] scholarly career path. It is possible, however, to see circumstances in which a

researcher could make a career out of reproducibility — if funders require it, for example, and particularly if funding is structured to provide compensation for the work. This would be the beginnings of a "reproducibility industry," following Stodden's example of NIH requirements creating a biostatistics industry. Then reproducibility research becomes an ongoing responsibility. The group also envisioned a situation in which principal investigators pay junior researchers, such as postdocs, to perform required reproducibility analyses.

There was a very vigorous conversation. Some of the group felt that there might be a place in industry to do this kind of work.

One point of view regarded the costs of curation and confirming reproducibility as a sort of research tax. If the loop is opened, and industry feeds back into the system, by supporting the work and being a well-behaved taxpayer, the system benefits. If, however, industry takes information and artifacts out of the system, and does not play by the rules and return support for the effort, then they become parasites. This is a point about the entrepreneurial state, and the role of the state in funding innovation, and how that percolates through into industry and back into society as a whole.

One participant noted that a few hardcore cynics in the room felt that if industry were to have a stake in keeping information or analysis secret, they would do so, despite any moral arguments. The group did not reconcile these views.

Another participant commented that it would be possible to change licensing terms for industry, allowing free access for educational purposes, and requiring paid access for commercial use.

One point of view regarded the costs of curation and confirming reproducibility as a sort of research tax. If the loop is opened, and industry feeds back into the system, by supporting the work and being a well-behaved taxpayer, the system benefits. If, however, industry takes information and artifacts out of the system, and does not play by the rules and return support for the effort, then they become parasites.

Moving on to the second question, the group noted that the negative phrasing of "How can a lack of reproducibility be prevented from causing legal issues for the publisher?" created confusion. The group restated the question as, "How can reproducibility cause legal issues for the publisher, and what legal implications or intellectual property issues might arise from artifact sharing?"

The discussion required some additional clarification, since the extent of the publisher's contribution — certifying the data and artifacts versus merely confirming that they exist — would affect the legal exposure. This is yet another of those one-size-does-not-fit-all situations, because exposure varies by jurisdiction. The European Union tends to be very strict about issues of data privacy, while the U.S. is a little more permissive. Differences in compliance and regulation will have an effect.

There are ethical as well as legal implications to consider. What is the publisher's relation to the artifact, and how do publishers frame their accountability? A publisher who says, "This is ours," had different obligations from a publisher who says, "We are linking to this on somebody else's website because we perceive it as relevant to your article."

There was another vigorous discussion, this one about non-reproducible work. The group agreed that reproducibility should not be expected for every data set and every bit of work. Some kinds of studies, such as observational studies, are inherently non-reproducible and should not be subjected to reproducibility verification. Other work might not be important enough to reproduce. Non-reproducible work will continue to be published. There are questions, though, about the other kind of non-reproducibility, in which another researcher tries to reproduce the experiment, the code, the data, and cannot. What is the publisher's obligation when this failure is reported? Will a paper on the failure be published? If not, is the publisher obliged to bring the failure to the attention of the organization that funded the work? What are the implications?

The group agreed that reproducibility should not be expected for every data set and every bit of work. Some kinds of studies, such as observational studies, are inherently non-reproducible and should not be subjected to reproducibility verification.

Finally, the group discussed the role of standards. Jurisdictions and funding-agency expectations differ greatly, making life very complicated for those trying to "do the right thing." If the various stakeholders' expectations could converge, at least somewhat, and move in the direction of a consistency, it would benefit everyone in the ecosystem.

A participant commented approvingly on the way the group divided non-reproducibility into several categories. Clearly, for example, physicists working on the big bang cannot rewind the universe to rerun the experiment. They are going to take measurements and make observations. As a first step, it might be useful to concentrate on reproducing the computational aspects of research. And we can't rewind the universe back and rerun the big bang. We've got what we've got, so one of the virtues of the discussion is a focus on the computational aspects as really the first step in reproducibility. A survey of human subjects might be expensive, or even impossible, to redo, if only because time has moved on. One can, however, reproduce the statistical tests and analyses, and confirming the validity of these low-hanging computational fruits is still very useful.

How will reproducibility be funded? What are the biggest challenges facing publishers? (Report from Group D)

The questions for discussion:

- *As the current scholarly publishing business model undergoes pressure from the tilt toward open access, and library budgets are further reduced, how will the added step of reproducibility be funded?*
- *What do you think are the biggest challenges facing publishers of scholarly engineering research on the reproducibility of research front?*
- *Other — please identify any other items we should consider in the future.*

The group started by addressing the elephant in the room: how will the activities discussed for a day-and-a-half be funded? They divided the question into two parts: economics and resources.

The economic issues also have two elements: additional funds, from grant set-asides or requirements (as Stodden and others suggested), and efficiency savings, whether by publishers or from libraries (as Ammerman and Toler described).

That we have a zero sum game here was a strong resonant theme — in the library budget, or the funder budget. The group recognized a need to start redefining the processes.

With respect to resourcing reproducibility approaches, the group discussed the need for process changes, the need for infrastructure, and potentially also the need for additional people and activities. The group also noted the discussions that automation might be a solution to scarce human resources, but stumbled trying to describe how that would happen in practice.

The group tried to elucidate how this activity would be funded. To help decide who might pay for these activities, they asked, "Who benefits?" Authors benefit from visibility, impact, citations, and tenure. Users benefit from better code and data, which elevate the foundations of their own work. Funders have better success outcomes, and higher efficiency in the project investments.

When the group moved on to consider who might pay, it struggled to find viable, sustainable business models through which the parties who benefit might shoulder some of the costs. The group touched on the negative feedback loop process mentioned earlier in the meeting: that improved quality and the greater effort needed to attain it might stem the flow of published output, whether articles or other first-class objects.

In light of this discussion, the answer to the second question ("What are the biggest challenges facing publishers …") would seem to be: "Figuring out what the business models actually will be in the future." There seemed to be a fundamental problem describing how to journey from the present state to the various visions of the future adduced earlier in the workshop.

While there might be a persisting role for a high-quality descriptive article format, we might actually be moving to a workflow that consists more of objects than articles, of certifiers rather than journals, and perhaps even of automation rather than reviewing.

In a zero-sum game, what might need to disappear in order to make room for improving the outputs, the outcomes, and the process? While there might be a persisting role for a high-quality descriptive article format, we might actually be moving to a workflow that consists more of objects than articles, of certifiers rather than journals, and perhaps even of automation rather than reviewing.

The group noted other, fairly obvious challenges to socialization and adoption of best practices in this area (c.f., "The PI Manifesto"). The group asked whether other communities have already solved some of challenges and distilled best practices for, say, software deposit, data citation, etc. — practices that this community could simply adopt.

There was a long discussion about the sustainability of the reproducible research workflow within the existing paper-centric, article-centric workflow — echoing again the challenges of a zero-sum game in a time of fixed funding.

The penultimate consideration of the group was the funding flow, particularly the 60% that is absorbed into institutional overheads. Perhaps reproducibility set-asides and carve-outs should come out of this pool, rather than pressuring the researcher-library-publisher "iron triangle" described by Ammerman.

Finally, amid the Cassandra-like prophesying, the group wanted to note that there is a ray of hope: There are near-term and immediate actions that stakeholders can take, actions that will deliver incremental and organic changes, experiments whose success may show the path forward.

One participant commented on the notion of the zero-sum game, referring to Toler's comments that we haven't yet fully embraced the web and realized its possible efficiencies. If that is so, then is this truly a zero-sum game? If we do embrace the web and alter how we do things, would we find savings without compensating losses?

Forster noted that there do indeed appear to be possible efficiency savings, not only from incremental changes to processes and workflows, but also from "completely turn[ing] them upside down as well." It's incumbent on the publishing community and the researcher and the institutional communities to figure out, within the things that they control, within their span of influence, what falls into both the incremental and the revolutionary efficiency savings camps.

Takeaway Messages

Before the close of the first day of the workshop, moderator John Keaton recapped the key discussion points. He noted that the group had considered the variability of code and software and data sets, the importance of working with smaller groups, motivating smaller communities that might want to do some of the reproducibility work. Terminology was an issue. There was also discussion of the scalability of the peer review process and the burden that expanded artifact review would put on reviewers. The need for standards was a recurring theme. Experiments and pilot projects were proposed as follow-up activities. Keaton invited participants to summarize their takeaways from the first day, and their paraphrased responses are collected below. Many themes were revisited in the panel and breakouts on the second day.

1. Regarding Research Reproducibility and Open Science

Differentiating Data and Software

One does not see very explicit differentiation between the needs of data and the software. These are very different species of animals. Evaluating them requires different communities, different vettings, and really in-depth appreciations of their remarkable differences. Software and data are related, and are both part of very massive digital transformation, but each has its own distinct characteristics.

A second observation is that we probably need to invite more participation by industry into these conversations, especially from some of the major corporations. They hold massive data sets, big data, and they have some pretty valuable experience. They, too, should contribute to the solutions we're trying to develop.

The Meaning of Reproducibility

There is wide variation in what we mean by reproducibility, as we move from one field to another and even within a single field. Enhancing reproducibility will bring significant benefits to the community, but it can require new or additional resources of time, effort, and finances. We may even need new types of organizations to maintain or correct the data we collect. There could be some open legal issues — privacy issues concerning the people from whom we collect the data and, particularly regarding software, intellectual property issues. Finally, there could be a need for some standardization to minimize variability and to make sure that the review process and the process of curation is not overly burdensome.

2. Rapidly Evolving Concepts of Research Curation

Tools and Metrics to Increase Value

It appears that the community (or some communities) lack a certain willingness to participate in building standards. The community wants standards that work, and means for sharing data and software that increase their value. But we have difficulty specifying what advantage the researcher gains by doing the necessary work. We talk about standards, about accessing and curating data, but may lack the metrics to measure the value of these actions. What tools will allow the research communities to find more ways of giving credit to everyone who participates in moving toward more reproducible research?

More Communication and Outreach

Those who are already working on various data citation and data metric standards (many of whom have bioinformatics rather than engineering backgrounds) need to do a better job of communicating what they've accomplished to wider communities. Better outreach would reduce the risk of redundancy and wasted time.

Think about Incentives

The publisher's role has been to help researchers get all of their research outputs through the publication process, and to integrate publication with their other activities to make it easier for them. Why are we using terms like "supporting evidence"? That's just material that typesetters don't know what to do with, so we post it as files and call it "supporting evidence." It's all research and it's all important, whether it makes it into the manuscript or not.

Today it became clear that it takes different sets of skills to validate code and data, and different sets of incentives to encourage people to do it. It won't just plug right into the authoring and peer-review system that we have now. We have to think through incentives to get people to do this, and how the mechanisms are going to work. That was a big insight.

Data vs. Description

A couple of things occur to someone new to this field: First, as research becomes increasingly computational, multi-disciplinary, and (often) global in scope, the relationship is likely to change between code and data on the one hand, and the related printed description of algorithms on the other. As code and data become more and more prevalent, will other researchers at some point begin to place greater value on the code and data than on the description of algorithms? Second, for this ecosystem to evolve, the tools have to be much more author-friendly. The more hurdles we remove from the paths of authors trying to upload their code and data, the more help they can get in preparing to upload code and data, the more likely they are to do it. Standards and the ecosystem will have to evolve to make this possible.

Beyond Content Hosting

To a member of an academic institution, it does appear that we need to develop better methods at all levels for rewarding, assessing, and encouraging contributions to reproducible research. We need a model different from the current system of publications and citations, which encourage researchers to keep data to themselves. The multidisciplinary aspects we have been talking about might encourage researchers to favor more multidisciplinary repositories when it comes time to publish.

More people will publish, in more nontraditional publications, venues, and repositories to get their work out there and used. We need to think about how to validate content, about standardization, about how publishers can do more than just host content, to add credibility or mine the data. Publishers need to move to servicing content, helping to mine it, rather than just hosting it.

Give Credit

The takeaway is that it's important to give credit for showing that reproducible code or well-described data is stable. That gives urgency to the need for allowing publication parts to morph over time, as new elements or versions or badges are added. We need ways to usefully store and point to code and data. Peer review might be after publication for good chosen artifacts. Finding

reviewers for code and data is a challenge, but grad students and postdocs might help and might actually enjoy it. Processes, using the term broadly, can also be validated, and approaches need to be field-specific. NSF and NIH already offer products like code and data so researchers can get credit; this might morph over time as we look at how to validate components.

Settle on Terminology

Though they have been covered, there are two points to repeat: We need to coalesce on terminology, and pursue the idea of identifying or badging content to let the public know at least whether it has been vetted or not, while not necessarily investing a lot of review resources. It's important to think about some tools or services to help authors submit and reviewers review. Also, we should think a little more about making code and data available even without an accompanying article, supported only by metadata.

Integrate Research and Publication

Thinking about reproducible science suggests the need to evolve a closer integration between the research activity and the publication activity. Putting it another way, we need a different model, one that integrates the systems and tools that support a very dynamic research process with the functions that a fixed publication process brings: curation, registration, preservation, certification of a set of results. These two systems are not integrated at all, and they must be, to have reproducible science.

Metadata, Metadata, Metadata

The first thing is: metadata, metadata, metadata. This is something that publishers could certainly get involved in. Regardless of all other issues, chewing into metadata — establishing structure, process, tools and other assistance — is the kind of bread-and-butter area where a lot of progress could be made. The reproducibility process would benefit, and many other processes would benefit as well.

Three other points recurred throughout the day: When we talk about releasing data, for example, there's a notion about gates for access. This might be worrying for some people, but it's reasonable to ask, "Who are you and why do you want to use this stuff?" This bears thinking about. Look at WikiLeaks. Five years ago, WikiLeaks was supposed to bring openness and transparency and sunlight. And look what it's doing now. Once it's available, the information won't always be used as intended, by good actors with good intentions.

Next, what will be the ultimate outcome of the processes we're talking about? Do they get us to a better representation of reality — faster, cheaper, and more efficiently? What does that mean for the community working on it? There's a finite amount of money, so we could be talking about fewer people doing research, albeit more reproducible research. That's just one potential future.

Finally, we're all doing a lot of analog thinking in a digital world, because it's hard to transit out from where we are right now to whatever the future might be. We keep coming back to the current model of what the research is, while many solutions might require completely reconfiguring the way this thing works.

Scaled Approach to Peer Review

From a software engineering perspective, one of the takeaways is the need to positively support development of better software in research phases. It's often a battle, with researchers arguing,

"Oh, this is just research code. It doesn't have to look that great." Having standards for reproducibility would let us say, "No, I think the research code needs to be a little cleaner than it is right now." With regard to peer review: a scaled approach seems most suitable. It's not clear yet what this would look like, but one-size-fits-all peer review, done the same way for every piece of code, probably won't work.

3. Sustainability: Creation, Peer Review, Curation

Needed: Sustainable Integration and Workflow

The group is taking seriously issues around peer review and changing publication models. Integration between systems is an issue. We're hearing that our current workflow is not sustainable. If so, can new, more sustainable models of integration and workflow be developed, within existing budgets and without a huge influx of additional funding?

Reform Tenure and Promotion Processes

We've talked a lot about how funders assess research proposals and make their choices, and about how they need to do a better job of recognizing contributions like the production of data sets for communities. That's all true, but the funders have done pretty well in this area, by and large. The really hidebound piece of this is the institutional tenure and promotion processes. These are substantially different from the processes the funders use. That's very important to remember. In particular, institutional tenure and promotion processes tend to squeeze out disciplinary differences in very nasty ways, because they draw from the entire campus rather than build processes specifically appropriate to the discipline the candidate works in.

Second, the discussion underscores the importance of separating questions of replicability and reproducibility from questions of open data and open-source software. Certainly, it is easier for a reviewer to demonstrate reproducibility or replicability if all the material is open and publicly available. But there are many, many scenarios in which — for data confidentiality, proprietary issues, industrial participation, or whatever reason — that is not possible. The ACM is a place where a lot of industry and academia meet and overlap. Its experiences teach important lessons about accommodating these relationships in our thinking about replicability and reproducibility. It's an oversimplification to call open data or open source or open everything the universal solution. That's just not going to be possible in very large sectors of research.

Third, back in the middle of the 20th century, we began fetishizing this notion of peer review. This had become a barrier to effective scholarly communication until the emergence of public preprint archives — things like the Los Alamos — now Cornell — arXiv. Also, frankly, a profligate waste of ways of human resource: this unmonetized reviewing time is just an enormous tax on the research community.

We need to be parsimonious in our applications of attestations of reproducibility and replicability. It makes sense to do this sparingly. And, in most cases, it makes sense to do it post-publication and separate from publication. It would be a very, very bad thing to build expensive and slow replicability processes into the publication system itself, or to set them up as things that authors are always motivated to seek because they confer extra credibility.

How to Manage Expanded Peer Review?

A unifying theme is: How in the world are we going to manage expanded notions of peer review? One out of every two or three participants has touched on this, and making peer review of artifacts too onerous will just impede scientific progress. This has to be a continuing discussion. We're not thinking that this is a showstopper. The important thing is that we feel collectively that we can discuss this going forward.

Another [emerging] point is that as nontraditional research artifacts become more important and more numerous, they may become part of the flawed tenure and promotion process. Research universities in the U.S. certainly pay more attention to intellectual property than they did 20 years ago. There may be tensions, perhaps, if we don't continue to have open and frank discussions within the research community about how truly open should be access to software, especially the products of publicly funded research. We want to respect the notion of proprietary rights as we look for enhanced reproducibility. There's also the question of where it's legitimate to make things proprietary and where things really should be open.

Invest Reviewer Resources Well

It's impressive that so many similar ideas have emerged in answers to different questions from different groups of people from different backgrounds. That says a lot about the opportunity we have to move forward in a positive way. One question that should be carefully examined is, "What is worthy of investing reviewing resources?" We seem to be starting to recognize the need for a different approach to determining what is worthy of these very, very precious resources. How do we allocate them and how do we make sure that the material outside of these decisions gets some type of treatment? This might be part of a broader discussion about peer review.

Balancing Business Models and Resources

There appears to be a tension between publishing's emerging business models, which seem to be more and more volume-driven, and the pressure that that places on the community's stressed reviewer resources, already at the breaking point.

Build Automated Tools

One of the take-home messages was from the ACM: some communities are already starting to address reproducibility fairly well. The challenge we face as an entire research community is to expand that to a much broader scale. To do that, we need automated tools for authors and reviewers. And resources. I'm not sure that postdocs and graduate students will be sufficient. In the US, most postdocs and graduate students now come from overseas. Many face fundamental communication challenges, even if they are able to do the technical work. What they write needs to be edited before it's sent out for review. That's a fundamental challenge.

A Threshold of Reproducibility

Any research article should satisfy basic standards of reproducibility. We have not discussed what has happened to the threshold of reproducibility we once had for articles. As research has gotten more complicated, incentives have shifted away from the completeness needed for reproducibility, to shorter page lengths, shorter descriptions, and rates of download and citation. Fundamentally, any published research article should be reproducible by a researcher who is intelligent and knows the field.

Cultural Challenges

The thing that jumps out is how much the enthusiasm for ensuring reproducibility varies among technical disciplines and subdisciplines — even among those represented in this room. Approaches and definitions differ. A number of colleagues have noted the need for standardization and structure, but this variation is a challenge that must be addressed.

The other thing that jumped out is the cultural challenge. Reviewers in each discipline do reviews in a certain way. If reproducibility review is to expand, at least into some subdisciplines, we might need to evolve how we understand how we review and think about the different kinds of reviewers who bring specialized expertise to the evaluation. It may be a cultural issue, or it may be an institutional issue, but the tenure and promotion process is a challenge that must be addressed.

Sharing Requires Trust (and New Sensibilities Regarding Reuse of the Work of Others)

The key issue can be summarized in one word: trust. Sharing requires trust. Publishing is, in a way, a trust business; readers want to know which information they can trust. Peer review is about creating trust. We haven't talked about certification enough today, but that is also a matter of trust. However we want this to evolve, we have to make sure that the trust factor is solidly in place.

Beware of Unintended Consequences

There can be unintended consequences. On one hand, we need to automate capture and bundling techniques. But many of these are predicated on standard platforms, whether hardware or software platforms. What is the unintended consequence of that? How might that urge to create replicable processes lend itself to stifling innovation, because it will be simply too cumbersome or difficult to figure out how to do appropriate peer review.

The second point refers to the comment that building peer review tools and infrastructure to confirm reproducibility will be expensive. What is our metric for reproducibility? How can we measure what we have now and what we will build? If there is not an existing baseline, we need to think about how we establish one. We're going to want a more direct and comprehensive measure of success, rather than such indirect measures as number of hits or a kind of impact factor.

Sustainability Is Key

Our workgroup reported that doing some pilots would make sense, but that needs to be refined. Many groups — the ACM, some geosciences groups, and others — have conducted reproducibility pilots, and demonstrated that it is possible. We need to be able to determine whether reproducible research can be established only with continuing heroic effort, or become the norm, sustainable and at the right scale.

To pick a subdiscipline and say, "Go, do a reproducibility pilot," is to tell several hundred people that they have to do something and do it quickly. It is not the same as asking whether they can ever make reproducibility work. A related point is a chicken-and-egg question. We've talked about the different definitions of reproducibility and related terms. What does assuring reproducibility really mean — providing a full infrastructure, or just documentation? Not defining our goal stops everybody from being successful — both researchers and the nascent reproducibility industry (to the extent that there is one). Until there is agreement on the problem and the solution, everyone is experimenting with different ideas, not building a system.

4. Immediate Next Steps for Research Curation and Peer Review

Inventory the Current Initiatives

There were two primary takeaways. One (and this often happens with multi-organizational meetings) is that we think about new ways to tackle a problem, but we don't always stop to do an assessment of what already exists. What groups are already working to solve some of these challenges? Perhaps they haven't been funded appropriately, or they are underpromoted. Reproducibility badging is a good example: There are lots of programs. How do we start to unify those? How do we bring those efforts together so that they become part of a common standard rather than five competing options?

Second, a lot of what was discussed requires behavior change. Changing behavior requires sensitization, but it also requires making the change easy. With the right incentives and the right mechanisms — from the publishers and funders especially — we're more likely to be successful

Avoid Duplicating Efforts

There is a danger of duplicating efforts from the standards perspective. Those not involved already should pay close attention to what RDA and groups like CODATA are doing. At International Data Week (Denver, CO, 11-17 September 2016) there was clearly a large number of people who are already thinking about this. As we try to see how these issues relate to individual publishers, we also need to look at them from the viewpoints of different disciplines, and see how they vary.

We're approaching reproducibility as a cross-disciplinary challenge. It is, but publishers have certain strengths that should be used: they can consolidate efforts and provide information in digestible form. Can they extend these strengths to reproducibility? We should discuss further where publishers fit into the conversation, and what solutions they might offer.

Is Special Consideration Needed for Industry?

First, it's exciting to see how many publishers are involved in this effort. Second, it is important to address the problem of integrating systems and new ways of supporting researchers — from initial innovation all the way through review, publication, and post-publication access. Third is an issue that hasn't been mentioned much: industrial research. In many areas of computer science, industry plays a significant part, and they face real proprietary and legal challenges to participating. We have to consider how to give industry avenues for participation.

Start Now, Perfect It Later

This is much more of a people problem than it is a technological problem. As a group, we need to socialize this evolution toward reproducible research. A phrase that resonated was "from adoption to acceptance to perfection." Get something out there, get it going, and then perfect it later — but let's start socializing this.

It is amazing that there actually are several serious efforts underway to do reproducible research, and give researchers credit for doing reproducible research. These include offering multiple kinds of publications to allow publishing artifact reviews, and badging. We need to build on that and find ways to integrate these things.

A Modern Conversation

First, this is not a conversation that we could have had two or five years ago. This is really a very modern conversation. There is a very dramatic shift in interest in and acceptance of reproducible research.

The breakout groups are reporting common threads and themes: we seem to be coalescing around some steps and some ideas. This is enormous progress. One take-home message is the power that intellectual property has to shape how this discussion proceeds, what programs are implemented in the community, and what infrastructure is built. Once these are set, it will be very difficult to change course — so supporting platforms, code, and the underlying software infrastructure are important.

We do need a principled approach to unify expectations and development. But we also should "just try stuff." Let's not worry whether it's the globally right thing, but take some initial steps and be ready to iterate and change and learn. It's a very new area and a complex collective-action problem. Whoever "we" is, however one identifies as one of the stakeholders, we can take some steps. The idea is to do something, and then have conversations just like this one, with different stakeholders coming together and bouncing ideas off one another. This is fundamental. This is how we're going to make progress on this question.

Make It Easy for Researchers

To repeat what almost everybody has said, though: There should not be duplication of effort. There is a lot of effort going into reproducibility, and there needs to be collaboration among publishers who have not been much involved in this space so far. Bringing them into standards creation is the key point if we want to bring reproducibility to fruition.

In terms of incentives, people have already said that we need to change community thinking about reproducibility. Tenure and promotion has been brought up a few times; the other side of the coin is to lower the barrier so that researchers can submit whatever they need to — to make it easy to do and to make whatever is uploaded useful to others. This should be very, very, very easy to do. The last point: While we have all been thinking strategically, we really do need to start somewhere. We need to show that we are serious with some pilot projects to start from, and build from the bottom up instead of the top down.

A Central Communications Hub Is Needed

To echo the comments of others: first, let's just do it. What we do will be very different in different communities. Our various communities need to address this. But what is missing, and what one would like to see come out of this meeting and perhaps out of IEEE, is some central place where we can gather information on this topic.

There are already definitions for the lexicon of reproducibility, but many people, even in this room, aren't aware of what they are. Many different groups have experience in replication. If there were to be some central way of communicating that information to other communities addressing reproducibility, that would be a great outcome of this meeting.

A Growing List of Stakeholders

When our workgroup started listing the numbers of stakeholders in the research reproducibility and review process, it just kept growing. It's not a new observation, but it did drive home how

many different groups have skin in this game, and how much needs to be done to align their interests to get everybody moving toward a common solution.

What Lies Ahead

It was very exciting to hear that so many people care about reproducibility. The main takeaway will actually be what happens tomorrow. How do we take this workshop and make sure that when we meet next year (and one hopes that everybody will meet next year) we can say, "We kicked this meeting off last year, and now we have so much more"? Perhaps this will be creating groups of volunteers who care about the issue. Perhaps not all of us will participate, but some likely will. What will be the framework for making this happen? Could the organizers enable that? It will be a very important as an outcome.

5. NSF and the Evolution of Concepts of Research and Curation

Limits of Funding

As you think about implementing these programs, remember that the funding pie is not getting any bigger. The struggle is that if the funder pays for programs like this, it's not funding research. As you think about implementation, then, remember to look, too, for business models that will work in the long term.

The other issue is the "coin of the realm" and the importance of thinking outside the box. Every four or five years, each NSF division brings in what we call a "committee of visitors." They come in and review processes and then give recommendations. During the Committee of Visitors' review of the Astronomy Division in 2014, there was a somewhat shocking revelation that, because application success rates have gotten so low, some younger faculty are now actually submitting the reviews of their proposals as part of their tenure packages. That has become a coin of the realm in some places.

We had no idea that that information was being used that way. The reviewers didn't know it. Nonetheless, the grant review had become a coin of the realm. There may be other things that NSF does that could also could be used as coins of the realm, though we just haven't realized it yet.

Build Vocabulary

The takeaway simply is that reproducibility of research is a very broad, complex issue, and our contribution, this workshop's contribution, to the problem can be perhaps to establish a vocabulary around the issues, so that we can settle on broad definitions and drive the discourse going forward.

Valuable Input for NSF

NSF staff are here to listen. In the Polar Programs Division, we take replication, reproducibility, and curation very seriously. The division has spent the past year working out new data management plans. To some extent, no matter how seriously we take it, we're still struggling to figure out what these definitions are. They are still very vague, in some sense. Workshops like this do offer a real chance to take advantage of the current window of opportunity to focus on this issue. A good fraction of the community is represented here, and there are program managers at NSF who really are looking for some direction from the community.

Conclusion and Next Steps

In the eyes of many, research reproducibility and open science are two sides of the same coin. The premise is that if everyone has access to the data that underlies a research undertaking, then the results and conclusions are more likely to be reproducible. Essentially the same argument applies to software. If a product of the research is code, it is the sharing of this code that provides most of the value to the community. While sharing data and code by no means guarantees reproducibility, it certainly opens new channels for peer validation. A very broad summary of remarks by workshop participants is that open science — and what it implies for shared code and data — may involve new burdens on researchers unless protocols for sharing are carefully crafted and respected. Beyond protocols for sharing, a dominant theme of the workshop was the challenge of making open science financially sustainable. None of the participants — whether from the ranks of the publishing houses that were represented, the government agencies, or individual researchers — felt that a clear path to sustainability was at hand.

Much of the discussion at the workshop involved the *manuscript versus supporting materials* conundrum. Probably the greatest consensus among all participants was that going forward, much of what has heretofore been referred to as "supporting material" will be viewed as having equal standing with the published manuscripts being supported. A continuing question is how to allocate scarce peer review resources between manuscripts and related supporting items, such as experimental protocols, data sets, and code. An example of how this question is now being answered successfully may be found in areas of computer science.

The computer algorithm FREAK appeared in a paper, "Fast REtinA Keypoint", published in the 2012 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, DOI: 10.1109/CVPR.2012.6247715. The paper has been cited 1,190 times according to Google Scholar. By this measure, the work is of considerable interest to the peer community, but the real value of the algorithm derives from its availability and its incorporation into the OpenCV computer vision software library where it is available for reuse in computer code by researchers worldwide. The value of the conference paper is in giving credit to the algorithm's creators, but the value of the algorithm itself is the extent to which it can be reused.

There are many such examples, and while peer review remains an important value-add by publishers, the true value of the research is only realized after reuse has occurred. To some extent, this offers insight into the complexity of reviewing nontraditional research artifacts. In the case of software, an initial peer review to determine whether the claimed effectiveness of an algorithm is plausible must be followed by use cases that occur only after the initial paper has been published. Many independent conversations at the workshop arrived at agreement on this point.

One of the recurring themes in presentations and discussions at the workshop was how to meet the challenge of the possible impermanence of the research archive. The movement toward open data has been embraced by the U.S. National Institutes of Health (NIH) in its launch of the Big Data to Knowledge (BD2K) program in 2014. Under the BD2K program, NIH will be launching a Data Commons Pilot to test ways to store, access, and share biomedical data and associated tools in the cloud. There are certainly significant costs that must be borne to make this sustainable, and the question arises as to how to allocate such costs. While there is no doubt that the U.S. government has the deep pockets necessary, there is possible cause for concern that long-term

financial commitments by the government may prove unreliable. A large open question as we conclude this report is "What are the alternatives in maintaining the research archives of the future?" Are distributed data archives the answer? Again turning to computer science, there are increasingly many examples of public data sets that have been created and archived by individual researchers and research teams (e.g., CIFAR [https://www.cs.toronto.edu/~kriz/cifar.html], MNIST [http://yann.lecun.com/exdb/mnist/], and ImageNet [http://image-net.org/about-overview] to name a few). The question arises as to what will happen to these when their creators turn their attention to other matters.

Deciding next steps remains a work in progress. Although there was fairly broad consensus on the challenges posed by (a) the curation of nontraditional research products, (b) the infrastructure needed to support open science, and (c) the overarching goal of advancing trust in science and research reproducibility, strategies for meeting the challenges are still in a formative stage. The workshop organizers are now considering topics on which further input from stakeholder communities should be sought. These include:

1. What steps can be taken to increase trust and enhance research reproducibility in all areas of science? Protocols for sharing data and code are important, but new approaches to traditional science publishing that include links to data and usable code are needed.

2. Where should funders and publishers concentrate resources in support of rapidly evolving forms of research curation? In view of the success of repositories and hosting services like GitHub, is the best policy frequently going to be one of benign neglect?

3. The sustainability question remains a matter of concern, not only as it relates to the archiving of nontraditional research products, but even as to how it can be achieved in the face of increasing demands for peer review needed to support the rapid proliferation of journals and conferences. Are there proxies and substitutes for traditional peer review — such as an increased reliance on research validation via post publication interest from the peer community and general public as measured by, say, citations, downloads, and the extent to which code and data are reused?

4. NSF and other funding agencies around the world are increasingly requiring "data management plans" as components of research proposals. As of this writing, content and format standards have not been established. Further discussion is needed to flesh out the requirements that are needed in such plans in order to ensure the greatest possible usefulness of research products. Research proposals should specify plans for all forms of research products and not be restricted to data alone.

5. Who should be at the table for the next workshop and the round of discussion it will support? Are there important stakeholders that were absent from the workshop reported in the above pages? While the workshop was intentionally focused on fields of interest to the IEEE, there are important activities being undertaken in other domains. The Data Commons initiative of NIH in the U.S. is a case in point. A follow-on workshop could benefit from perspectives on such efforts to enhance research reproducibility. Work is needed to inventory current initiatives across a broad cross-section of scientific research. While the workshop included participants from Europe, a possible future workshop could benefit from greater involvement of non-U.S. funding agencies.

These are many questions that remain and that can be further addressed if there is a follow-on workshop. The timing of such a workshop needs to be sorted out through dialogue among the

organizers, funders, and important stakeholders within the research community. In closing, it is worth noting that while there have been a good number of workshops supported by various agencies in both the U.S. and abroad, the workshop reported above was somewhat special in its having significant participation from publishers. Publishers have been an essential part of research curation and dissemination for centuries, and while their role has unquestionably changed in the age of the World Wide Web, their value in curating research and preserving knowledge remains high. They continue in their role of neutral arbiters of research, and as the custodians of unbiased peer review, this role has never been more important for ensuring research integrity in an age when science is all too frequently being publicly denigrated and questioned.

Appendix

Agenda

The First IEEE Workshop on

The Future of Research Curation and Research Reproducibility

At the Marriott Marquis, Washington, DC, USA, 5-6 November, 2016
National Science Foundation Award # 1641014
The NSF-sponsored IEEE Workshop on "The Future of Research Creation and Research Reproducibility" will explore the frontiers of research reproducibility in the engineering and mathematics domains. Specifically, the Workshop will address questions surrounding the content, review of the content, and the economic impact of reproducibility across the research and publishing ecosystem.

Agenda

Friday November 4, 2016
6:30 PM — 10:00 PM
Group Dinner – Kellari Taverna 1700 K St NW
All that would like to walk or cab share together can meet in the lobby at 6:00pm.

Saturday November 5

7:00 Breakfast (Supreme Court Room)

General Session (Capital/Congress)

8:00 Welcome
John Baillieul, Boston University

8:10 Amy Friedlander, Deputy Division Director, National Science Foundation

8:30 Clifford Lynch, Executive Director, Coalition for Networked Information
Overview of the reproducibility landscape

8:50 Plenary Panel — New forms of Content
What are the essential products of scholarly research, how will these be likely to change in the future and how can the results of the research be accurately reproduced? This panel will identify new types of content and the challenges of reproducibility.

Panel Moderator: Larry Hall, Dept. of Computer Science & Engineering, University of South Florida
Panelists: Jelena Kovačević, Carnegie Mellon; Simon Adar, Code Ocean; Eric Whyne, DataMachines.io; Sheila Morrissey, Portico

9:45 Break

10:00 Q&A — New forms of content

10:30 Breakout discussion groups (Cherry Blossom, Magnolia, Dogwood)

11:30 Group Report-outs, New Forms of Content (Capital Congress)

12:00 Lunch (Supreme Court Room)

1:00 Plenary Panel — Peer Review and Quality Assurance
As non-traditional types of research products (i.e., data and software) become a significant component of the curated research record, how should quality assurance be organized? Some questions to be pondered: Do we need to provide a common platform? Can we run experiments using different software and environments? How to address possibility of proprietary software (e.g. compilers).

Panel Moderator: Sheila Hemami, Director, Strategic Technical Opportunities, Draper
Panelists: Bernie Rous, ACM; Pierre Vandergheynst, EPFL; Jennifer Turgeon, Sandia National Labs; Eleonora Presani, Product Manager, Scopus, Elsevier

2:10 Q&A — Peer review and Quality Assurance

2:40 Break

3:00 Break out groups

4:00 Group report-outs, Peer Review

4:30 Day one summary

5:30 Adjourn, Day 1

6:30 PM — 10:00 PM
Group Dinner — Brasserie Beck 1101 K St NW
All that would like to walk or cab share together can meet in the lobby at 6:00pm.

Day 2 November 6, 2016
7:00 Breakfast (Supreme Court Room)

General Session (Capital/Congress)
8:00 Welcome Day 2 — Michael Forster, Managing Director, Publications, IEEE

8:10 Plenary Panel — Economics of reproducibility
As the current scholarly publishing business model undergoes pressure from the tilt toward open access, and library budgets are further reduced, how will the added step of reproducibility be funded? Panel will discuss funding scenarios.

Panel Moderator: Gianluca Setti, Dept. of Engineering, University of Ferrara, Italy
Panelists: Todd Toler, John Wiley & Sons; Jack Ammerman, Boston University; Victoria Stodden, UIUC; Dan Valen, Figshare

9:30 Q&A — Economics of reproducibility

10:00 Break

10:15 Break out groups

11:15 Group report outs — Economics

12:30 Lunch (Supreme Court Room)

1:15 Recap of the Workshop and next steps

2:00 Meeting Concludes

Workshop Objectives

This Workshop will provide a forum for constructive dialogue between publishing professionals and members of various stakeholder communities with a shared interest in public dissemination of scholarly research in engineering and the information sciences, as represented by a variety of IEEE societies including, Signal Processing, Control Systems, Robotics and Automation, Information Theory, and Circuits and Systems. Participants will explore three interrelated and increasingly important questions concerning future approaches to deriving the maximum possible benefit from the products of engineering research:

1. New forms of Content and Radically New Approaches to Content Creation
From the dawn of writing and movable type printing until very recently, the results of scholarly inquiry have been communicated through printed documents. Scholarly articles comprising archival journals have been the medium though which the work of researchers has become known and used to enable further research. Time-honored practices in publishing are currently undergoing tumultuous disruption on a number of fronts. Beginning with the appearance of the World Wide Web researchers began self-archiving preprints and even copies of papers that had been published and for which publishers held copyrights. After the appearance of digital archives of downloadable pdf versions of published articles (IEEE Xplore, Elsevier Science Direct, AIP Scitation, etc.), the proliferation of new research sharing models grew quickly to include increasingly sophisticated self-archiving, preprint servers (arXiv), and early university research repositories (DSpace). The landscape has continued to change with increasingly popular web-based collaboration tools that support not only collaborative writing but also code development, and data curation. A timely question is: What are the essential products of scholarly engineering research, how will these be likely to change in the future?

2. Peer Review and Quality Assurance of Curated Research
As nontraditional types of research products (e.g. experimental protocols, data, and software) become increasingly significant components of the curated engineering research record, how should quality assurance be organized and paid for? As research becomes increasingly "versioned", how will peer review be applied when the version of record of published research is subject to continuous revision and updating? With virtually every product of research — from papers to software to data sets themselves — being updated on a continuing basis, what new forms of peer validation will be needed. Will it be possible to have persistent links between published papers and supporting software, experimental records, and data?

3. What are economically sustainable approaches to providing public access to engineering research products?
The goal of ensuring that future engineering research will be maximally reproducible underlies all these questions. Workshop presentations by publishing professionals will explore current and planned approaches to data and software curation in engineering and other disciplines. There will also be presentations by data professionals who currently provide platforms for such curation as well as those engaged in research on fundamental data science, data infrastructure, and cyber-infrastructure. Using new curation and publishing technologies to most effectively harvest value from curated research products will be explored in alignment with stated National Science Foundation objectives of developing new advances in data infrastructure and analytics, reproducibility, privacy and protection, and research in the human-data interface.

The topics to be covered by the Workshop include:
- Data curation -ethical data management
- Software curation
- Versioning of archival literature
- Research reproducibility -including reproducibility metrics
- Peer review -data, software, versions - how to manage
- The evolving relationship between scholarly publishers, researchers, and research libraries

The participants will include researchers, representatives of selected publishers, data curation professionals, engineering

researchers, and public access representatives from the U.S. National Science Foundation and other U.S. Government research agencies. The preliminary date and venue of the Workshop are 5-6 November 2016 in the Washington, DC, area.

Steering Committee

Chair: John Baillieul, Boston University

Larry Hall, University of South Florida	José M.F. Moura, Carnegie Mellon
Sheila Hemami, Draper Labs	Gianluca Setti, University of Ferrara
Michael Forster, IEEE	Gerry Grenier, IEEE
Fran Zappulla, IEEE	John Keaton, IEEE

Background Reading

1. D. Goodyear, 2016. "The Stress Test: Rivalries, intrigue, and fraud in the world of stem-cell research, Annals of Science," *The New Yorker*, Feb. 29, 2016.
2. J. B. Buckheit and D. L. Donoho. (1995). "WaveLab and reproducible research," Dept. of Statistics, Stanford Univ., Tech. Rep. 474.
3. Turnbull, H.W. ed., 1959. The Correspondence of Isaac Newton: 1661-1675, Volume 1, London, UK: Published for the Royal Society at the University Press. p. 416.
4. Special Online Collection: Hwang et al. Controversy — Committee Report, Response, and Background; http://www.sciencemag.org/site/feature/misc/webfeat/hwang2005/
5. Begley, C. G. and L. M. Ellis, "Drug development: Raise standards for preclinical cancer research," *Nature* 483(7391): 531-533, 2012; http://www.nature.com/nature/journal/v483/n7391/full/483531a.html
6. Ioannidis, J. P. A., et al. "Repeatability of published microarray gene expression analyses." *Nat Genet* 41(2): 149-155, 2009; doi:10.1038/ng.295
7. "How Science goes Wrong," *The Economist*, 19 October 2013; http://www.economist.com/printedition/2013-10-19.
8. D. Donoho, A. Maleki, I. Rahman, M. Shahram, and V. Stodden. "Reproducible research in computational harmonic analysis," *Computing in Science & Engineering*, 11(1):8-18, Jan.-Feb. 2009.
9. http://sites.nationalacademies.org/DEPS/BMSA/DEPS_153236
10. Brian Nosek, "Improving and Rewarding Openness and Reproducibility," NSF Distinguished Lecture Series in Social, Behavioral and Economic Sciences, https://www.nsf.gov/events/event_summ.jsp?cntn_id=13526
11. http://www.acm.org/data-software-reproducibility
12. http://www.stm-assoc.org/events/innovations-seminar-2015/
13. NSF's Public Access Plan: Today's Data, Tomorrow's Discoveries, National Science Foundation, March 18, 2015, NSF 15-52. http://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf
14. Zhaodan Kong, Nathan Fuller, Shuai Wang, Kayhan Ozcimder, Erin Gillam, Diane Theriault, Margrit Betke, and John Baillieul, "Perceptual Modalities Guiding Bat Flight in a Native Habitat," *Scientific Reports*, Nature Publishing Group. http://www.nature.com/articles/srep27252
15. Flight data for cave emergence of Myotis velifer, July, 2013. John Baillieul, Curator. Bat Data
16. 2010-2013 New York City Taxi Data, Dan Work, Curator. Taxi Data.
17. FORCE11 Data Citation Workshop on February 2, 2016.
18. Workshop Series to Gauge Community Requirements for Public Access to Data from NSF-Funded Research, https://www.nsf.gov/awardsearch/showAward?AWD_ID=1457413&HistoricalAwards=false.
19. Data Management and Data Sharing Workshop for Science and Technology Studies. https://www.nsf.gov/awardsearch/showAward?AWD_ID=1430608&HistoricalAwards=false.
20. Increasing Access to Federal Court Data Workshop--Fall 2015. https://www.nsf.gov/awardsearch/showAward?AWD_ID=1551564&HistoricalAwards=false.

21. Support for Rise of Data in Materials Research Workshop; University of Maryland; June 29-30, 2015. https://www.nsf.gov/awardsearch/showAward?AWD_ID=1542923&HistoricalAwards=false.
22. Dear Colleague Letter: Citizen Science and Crowdsourcing - Public Participation in Engineering Research,https://www.nsf.gov/pubs/2016/nsf16059/nsf16059.jsp.
23. Center for Open Science https://cos.io/.
24. P. Basken, "Can Science's Reproducibility Crisis Be Reproduced?", *Chronicle of Higher Education*, March 03, 2016, on-line at Article.
25. V. Stodden, F. Leisch, R.D. Peng. *Implementing Reproducible Research*, 2014, Chapman and Hall/CRC, 448 Pages, ISBN 9781466561595.

Endnotes

[1] Holdren JP, 2013. "Memorandum for the Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research." Executive Office of the President, Office of Science and Technology Policy. https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[2] National Science Foundation, 2015. NSF's Public Access Plan: Today's Data, Tomorrow's Discoveries: Increasing Access to the results of research funded by the National Science Foundation. NSF 115-52. https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf

[3]Examples of revision projects include: A Method to Retrieve Non-Textual Data from Widespread Repositories (1450545/Hovy); Portable, Secure Emulation for Digital Preservation (1529415/Brown); Using search engines to track impact of unsung heroes of big data revolution, data creators (1565233/Godzik); Supporting Public Access to Supplemental Scholarly Products Generated from Grant Funded Research (1649703/Lehnert, Stodden, and Berman)

[4] M. Barni ; F. Perez-Gonzalez (2005). Pushing science into signal processing. IEEE Signal Processing Magazine (Volume: 22, Issue: 4, July 2005), pp 119-120. http://ieeexplore.ieee.org/document/1458324/

[5] J. Kovačević. How to encourage and publish reproducible research. In Proc. IEEE Int. Conf. Acoust., Speech Signal Process., pages IV:1273-1276, Honolulu, HI, Apr. 2007.

[6] P. Vandewalle, J. Kovačević and M. Vetterli. Reproducible research in signal processing: What, why and how. IEEE Signal Process. Mag., 26(3):37-47, May 2009.

[7] Wiles, Andrew (1995). "Modular elliptic curves and Fermat's Last Theorem" (PDF). Annals of Mathematics. 141 (3): 443–551. doi:10.2307/2118559. JSTOR 2118559. OCLC 37032255.

[8] Taylor R, Wiles A (1995). "Ring theoretic properties of certain Hecke algebras". Annals of Mathematics. 141 (3): 553–572. doi:10.2307/2118560. JSTOR 2118560. OCLC 37032255. Archived from the original on 27 November 2001.

[9] W. S. Hwang et al., Evidence of a Pluripotent Human Embryonic Stem Cell Line Derived from a Cloned Blastocyst, Science 303, 1669 (2004)

[10] Kennedy D (2006) "Retraction of Hwang et al., Science 308 (5729) 1777-1783." Science 311 (5759), pp. 335 DOI: 10.1126/science.1124926

[11] Vetterli M, Kovačević J, Goyal VK (2014). Foundations of Signal Processing, 3rd ed. Cambridge University Press. 738 pp. ISBN-13: 978-1107038608.

[12] https://www.apache.org/licenses/LICENSE-2.0

[13] Van Zee FG, van de Geijn RA (2015). BLIS: A framework for rapidly instantiating BLAS functionality. ACM Transactions on Mathematical Software, 41(3): Article 15. Association for Computing Machinery.

[14] http://jats.niso.org/1.1/

[15] http://json-ld.org/

[16] MIT Ad Hoc Task Force on the Future of Libraries (2016). Institute-wide Task Force on the Future of Libraries, https://www.pubpub.org/pub/future-of-libraries

[17] https://figshare.com/

[18] http://opendataenterprise.org/transition-report.html

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

[20] http://www.codata.org/

[21] https://databank.illinois.edu

[22] Fallaw C, Dunham E, et al. (2016) Overly honest data repository development. Code{4}lib Journal 1(34):2016-10-25. http://journal.code4lib.org/articles/11980

[23] Jubb M (2016). Review: Embedding cultures and incentives to support open research. A review commissioned by the Wellcome Trust focusing on cultures and incentives in open research, and mechanisms to address key

challenges. Figshare, https://figshare.com/articles/Review_Embedding_cultures_and_incentives_to_support_open_research/4055514

[24] Alberts B, Cicerone RJ, et al. (2015) Self-correction in science at work, Science 348(6242):1420-1422. DOI: 10.1126/science.aab3847. http://science.sciencemag.org/content/348/6242/1420.full

[25] Nosek BA, Alter G, et al. (2014) Transparency and Openness Promotion (TOP) Guidelines. Open Science Framework. https://osf.io/9f6gx/

[26] Silva L. (2014) PLOS' New Data Policy: Public Access to Data. PLOS http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/

[27] Howard J. (2013) Rise of 'Altmetrics' Revives Questions About How to Measure Impact of Research. Chronicle of Higher Education. 3 June 2013. http://www.chronicle.com/article/Rise-of-Altmetrics-Revives/139557/

[28] Barba LA (2016). The hard road to reproducibility. Science 354(6308): 142 DOI: 10.1126/science.354.6308.142. http://science.sciencemag.org/content/354/6308/142

[29] Chodacki J, Cruse P, et al. (2016) A Healthy Research Ecosystem: Diversity by Design, *The Winnower* 4:e146047.79215 (2016). DOI: 10.15200/winn.146047.79215. https://thewinnower.com/papers/4172-a-healthy-research-ecosystem-diversity-by-design

[30] Treadway J, Hahnel M., et al (2016). The State of Open Data Report: A selection of analyses and articles about open data, curated by Figshare. Foreword by Sir Nigel Shadbolt.