Workshop Report

NSF Workshop on Supporting Scientific Discovery through Norms and Practices for Software and Data Citation and Attribution

April 20, 2015

Report from the National Science Foundation- and Sloan Foundation-funded workshop held January 29-30, 2015 in Arlington, Virginia to address challenges in software and data citation that can be made actionable.

Written by:
Stan Ahalt, RENCI, University of North Carolina at Chapel Hill
Tom Carsey, Odum Institute, University of North Carolina at Chapel Hill
Alva Couch, Tufts University
Rick Hooper, Consortium of Universities for the Advancement of Hydrologic Science, Inc.
Luis Ibanez, Google Inc.
Ray Idaszak, RENCI, University of North Carolina at Chapel Hill
Matthew B. Jones, NCEAS, University of California, Santa Barbara
Jennifer Lin, Public Library of Science
Erin Robinson, Foundation for Earth Science

This material is based on work supported by the National Science Foundation under grant number 1448360 with additional support from the Alfred P. Sloan Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Sloan Foundation.

Table of Contents

Workshop Organizing Committee

Stanley C. Ahalt, Ph.D., PI, Meeting Chair
Director, Renaissance Computing Institute (RENCI)
PI, Water Science Software Institute NSF S2I2 Conceptualization Award
Professor, Department of Computer Science, UNC-Chapel Hill
Director of Biomedical Informatics Service, NC TraCS
DataNet Federation Consortium Facility Lead
Chair, National Consortium for Data Science Steering Committee
ahalt@renci.org

Thomas M. Carsey, Ph.D., Co-PI
Thomas J. Pearsall Distinguished Professor, Department of Political Science, UNC-Chapel Hill
Director, Odum Institute for Research in Social Science, UNC-Chapel Hill
Dataverse BIGDATA Co-PI
DataBridge BIGDATA Co-PI; iRODS-Dataverse integration for DataNet Federation Consortium
carsey@unc.edu

Alva Couch, Ph.D.
Associate Professor, Department of Computer Science, Tufts University
Director, CUAHSI Water Data Center
acouch@cuahsi.org

Rick Hooper, Ph.D.
Executive Director, Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI)
rhooper@cuahsi.org

Luis Ibanez, Ph.D.
Google, Inc.
luis.ibanez@gmail.com

Ray Idaszak
Director of Collaborative Environments
Renaissance Computing Institute (RENCI), UNC-Chapel Hill
rayi@renci.org

Matthew B. Jones
Director, Informatics Research and Development, National Center for Ecological Analysis and Synthesis (NCEAS)
PI, Institute for Sustainable Earth and Environmental Science NSF S2I2 Conceptualization Award (ISEES)
DataOne Leadership Team Member
jones@nceas.ucsb.edu

Jennifer Lin, Ph.D.
Senior Product Manager, Public Library of Science (PLOS)
jlin@plos.org

Erin Robinson
Executive Director, Foundation for Earth Science (FES), the organization that supports the Federation for Earth Science Information Partners (ESIP)
erinrobinson@esipfed.org

Executive Summary

Software is as essential as data in the modern practice of science. When scientists share with each other not only research results, but also data and software, it vastly amplifies the reach, relevance, and transparency of science. Yet there are substantial social, systemic, and technological barriers that prevent scientists from sharing data and software. Scientific researchers – particularly academics – are embedded in a reputation economy in which tenure, promotion, and acclaim are achieved through influential research results. Tenure and promotion decisions are typically blind to a researcher's contributions to shared data or software, despite the crucial role of these activities in the scientific endeavor. Compounding the problem, there are no standard practices for citing data and software, giving appropriate credit to contributors, or measuring the impact and value of data and software contributions. Although numerous data and software sharing repositories exist, each uses a slightly different approach and many scientists still distrust the public access model, preferring to share data and software only by personal request, which assures attribution through personal contact and implicit social contract but substantially limits the reach and benefit of shared data and software.

The research community urgently needs new practices and incentives to ensure data producers, software and tool developers, and data curators are credited for their contributions. This National Science Foundation (NSF)-sponsored workshop facilitated a national, interdisciplinary discussion and exploration of new norms and practices for software and data citation and attribution to inform the Software Infrastructure for Sustained Innovation (SI2) and Science of Science and Innovation Policy (SciSIP) NSF programs. Participants identified social and technical challenges facing current software development and data generation efforts and explored viable methods and metrics to support software and data attribution in the scientific research community. A consensus throughout the workshop was a strong sentiment that it is time to move beyond discussion of the issues and begin to establish pilot projects that endeavor to implement and experiment with actionable ideas. Section 3 presents a full listing of actionable plans discussed at the workshop; highlights of these include:

- Request that publishers and repositories interlink their platforms and processes so that article references and data set or software citations cross-reference each other.
- Request that the research community develop a primary consistent data and software citation record format to support data and software citation.
- Request that an organization (as yet unidentified) develop guidelines for trusted software repositories for science (similar to trusted digital data repositories).
- Ask federal funding agencies to require every Principal Investigator (PI) to have a permanent human identifier (e.g., ORCID, which resolves critical issues of identifying individuals).
- Data and software repository landing pages should describe the full provenance of the data using appropriate standards.
- Authors should be able to cite data and software in their articles at an appropriate level of granularity.
- Federal funding agencies should support an effort to convene key players to identify and harmonize standards on roles, attribution, value, and transitive credit (in an extensible framework). All key sponsors would be recognized.
- Agencies, publishers, societies, and foundations should fund implementation grants to identify and measure data and software impacts in a way that is relevant to stakeholders and research communities.

1. Overview

As part of its Software Infrastructure for Sustained Innovation (SI^2) program and Science of Science and Innovation Policy (SciSIP) program, the NSF is exploring new norms and practices for software and data citation and attribution in research communities so that data producers, software and tool developers, and data curators are credited for their contributions. To assist with this endeavor, the NSF issued a Dear Colleague Letter titled "Supporting Scientific Discovery through Norms and Practices for Software and Data Citation and Attribution" inviting collaborative workshop and exploratory research proposals aiming to address these issues.[1]

In response to this invitation, RENCI,[2] working with an organizing committee, planned and hosted a workshop to facilitate a national, interdisciplinary discussion and exploration of norms and practices for software and data citation and attribution, with the goal of informing the further development of the SciSIP and SI^2 programs. Workshop participants gathered January 29 and 30 at the Hilton Arlington Hotel in Arlington, Virginia to discuss the social and technical challenges facing current software development and data generation efforts and to explore viable methods and metrics to support software and data citation and attribution in scientific research communities. The workshop was one of three hosted by organizations across the United States focusing on a variety of subjects designed to inform the NSF SI^2 and SciSIP programs.

Among the 48 workshop attendees, 24 represented university departments and institutes, nine represented government agencies (including two international agencies), seven came from corporations, and eight represented non-profit organizations. Organizations represented at the workshop included, among others, the American Geophysical Union (AGU), the Commonwealth Scientific and Industrial Research Organization (CSIRO), the Consortium for the Advancement of Hydrologic Science, Inc. (CUAHSI), Elsevier, the Foundation for Earth Science (FES, the entity that supports the Federation for Earth Science Information Partners, or ESIP), Australian National Computational Infrastructure (NCI), Information International Associates (IIA), John Wiley & Sons, Inc., Mozilla Foundation, the National Institutes of Health (NIH), the National Aeronautics and Space Administration (NASA), the National Science Foundation (NSF), the National Oceanic and Atmospheric Administration (NOAA), the Public Library of Science (PLOS), the Research Data Alliance (RDA), SAGE Publishing, the Sloan Foundation, and the Woods Hole Oceanographic Institution. A list of attendees and their organizational affiliations is included in Appendix A.

This workshop facilitated a wide-ranging interdisciplinary discussion and exploration of new norms and practices for software and data citation and attribution. Participants identified social and technical challenges facing software development and data generation efforts and explored viable methods and metrics to support software and data attribution in research communities. The workshop emphasized actionable plans that will enable the broader research community to implement new software and data attribution practices.

[1] Dear Colleague Letter (14-059) - Supporting Scientific Discovery through Norms and Practices for Software and Data Citation and Attribution, April 11, 2014, http://www.nsf.gov/pubs/2014/nsf14059/nsf14059.jsp.

[2] Renaissance Computing Institute; http://www.renci.org/; an institute of the University of North Carolina at Chapel Hill.

2. Workshop Summary

Funding for the workshop was awarded by the NSF on September 1, 2014. The workshop organizing committee announced its plans for the workshop at https://softwaredatacitation.org/ and assembled a representative spectrum of attendees from a variety of research domains and sectors. The workshop website was setup as a wiki that organizers and participants used to plan the plenary, panel, and breakout presentations. The workshop was organized to be largely participant driven, a structure sometimes referred to as an "unconference" format.[3] To help inform the workshop discussions, a GitHub website[4] was published in advance of the workshop and all attendees were encouraged to submit and/or comment on use cases illustrating challenges in software and data citation. Each use case contained a full description, contributors, statement of goals, existing efforts to date, set of actionable outcomes, and other pertinent information.

Twenty-two use cases were collected prior to the workshop. Four use cases focused primarily on software citation challenges, 13 dealt primarily with data citation challenges, and five addressed both software and data citation challenges. The day before the workshop, members of the organizing committee arranged the 22 participant-contributed use cases into eight topics aligned on common themes. These themes formed the basis of the morning plenary session discussions and afternoon breakout sessions. In the participant-driven spirit of the unconference, two additional topics were added by participants during the Thursday morning plenary session. These 10 consolidated topics were written on flipchart pages and hung around the plenary meeting room. Topics were as follows:

- *Best Practices:* How can we create and promote best practices for data and software citation (abbreviated as "D/S citation" in this document)?
- *Beyond the University Model:* What are the challenges with the university model and reward structure with respect to D/S citation (and should we go beyond it)?
- *Computational Reproducibility:* How can we enable D/S citation for computational reproducibility?
- *Difference between Software and Data:* What are the differences between software and data with respect to citation?
- *Identify and Version:* How can we identify authors and version contributions for D/S citation?
- *Interoperable Frameworks:* How can we create interoperable frameworks for D/S citation?
- *Large and Complex Data:* How can we address special issues with large data sets and computational data?
- *Social Barriers:* What are some strategies for breaking down the social barriers associated with D/S citation?
- *Subsetted, Derived, and Aggregated:* How should we cite subsetted, derived, and aggregated data and software?
- *Useful Metrics:* How can we establish useful D/S metrics?

The first day of the workshop, January 29, began with plenary sessions in the morning followed by breakout sessions in the afternoon. Stan Ahalt, workshop Principle Investigator and host, spoke first to welcome the attendees, recognize the workshop sponsors, and describe the workshop's goals and desired outcomes. Following this, Stavros and Costa Michailidis from

[3] See http://en.wikipedia.org/wiki/Unconference.
[4] See https://softwaredatacitation.org/Pages/Use-Cases.aspx.

Know Innovation, Inc. explained their roles as professional workshop facilitators supported by a grant from the Sloan Foundation. During the rest of the morning plenary session, attendees heard from a series of speakers who were directly involved in data and software citation efforts. As they listened to the speakers and participated in discussions following each presentation, attendees were instructed to articulate, on post-it notes, possible actionable ideas that could help address key challenges identified by the speakers and attendees. Participants added their notes to the flipchart pages hung around the room representing the meeting's 10 major discussion topics. In turn, groups in each breakout session were instructed to consider the notes attached to their flip chart page and identify key actionable ideas for their topic. These served as the basis for developing formal action plans later in the workshop.

Plenary session presentations on the morning of the first day were as follows:

- Stan Ahalt (RENCI, PI)
  Welcome: Overview, Goals, and Desired Workshop Outcomes
- Luis Ibanez (Google, Inc., Featured Guest Speaker)
  Supporting Reproducible Scientific Research with Open Source Practices for Software and Data Citation and Attribution: A 15-Years Perspective and Vision for the Future
- Dan Katz (NSF)
  Metrics and Citation for Software (and Data)
- Thomas Carsey and Jonathan Crabtree (UNC Odum Institute for Research in Social Science)
  Automated Data Citation in the Social Sciences Using the Dataverse Network Open-Source Software
- Sweitze Roffel and Mike Taylor (Elsevier)
  Linking Data In and Outside a Scientific Publishing House – A Perspective from a Publisher
- Jennifer Lin (PLOS) and Matthew Jones (UCSB, NCEAS)
  Make Data Count: Open Source Software Collecting Metrics for Data and Software Use
- James Howison (University of Texas, Austin)
  How Software is Mentioned/Cited in the Biology Literature

The morning plenary generated many actionable ideas for each major discussion topic, as shown in Figure 1.

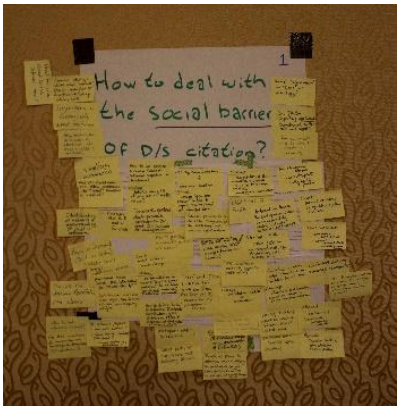

Figure 1: Representative flipchart for the Social Barriers topic.

The facilitators and organizing committee then organized the 10 discussion topics into a series of breakout sessions. Topics discussed during breakout sessions on the afternoon of January 29 were Social Barriers, Computational Reproducibility, Subsetted Derived and Aggregated, Useful Metrics, and Difference Between Software and Data. Topics discussed during breakout sessions on the morning of January 30 were Large and Complex Data, Identify and Version, Best Practices, Interoperable Frameworks, and Beyond the University Model. Attendees were invited to choose which breakout sessions to attend based on their interest and experience.

Participants in each breakout session were tasked with providing the following for their topic:

- A summary of the challenge,
- A summary of why it is important,
- A summary of why it is not yet solved, and
- Critical Actions (~3-5) needed to solve the challenge.

Each breakout session group had two hours to create slides that addressed the above four sub-topics. “Critical Actions” were defined as relevant, concrete, rational, aggressive, and understandable calls for action that would help to realize meaningful solutions. Breakout session participants were also asked to be cognizant of the information in the NSF Dear Colleague Letter that the workshop was designed to respond to.

Each group selected a moderator, note-taker, and presenter. Following breakout sessions on each day, all workshop attendees reconvened to share their group’s outcomes and hear from other breakout groups. See Section 3 for summaries of the outcomes for each breakout session. After each presentation, a “How, Wow, Now, Why” matrix was used to gather feedback from the attendees related to the impact and likelihood of success of each Critical Action. See Section 4 for a summary of these feedback sessions.

The workshop concluded at 12:30 p.m. Friday, January 30, with closing comments from workshop PI Stan Ahalt and NSF representative Daniel Katz. This was followed by a closed session among members of the workshop organizing committee to generate writing assignments, a timeline, and due dates.

An initial draft of the workshop report was created by March 20, 2015, announced, and made available to workshop attendees as an editable Google Document for open review and comment. The open review period lasted through March 30, with new versions created integrating community comments. The final draft was then reviewed for typographical errors and grammar, and submitted to the NSF and posted to the workshop wiki for public dissemination.

3. Summary of Critical Actions

This section summarizes the Critical Actions for each of the 10 breakout-session topics. “Critical Actions” are defined as relevant, concrete, rational, aggressive, and understandable calls for action that would help to realize meaningful solutions.

3.1 Best Practices

This session addressed the question: How can we create and promote best practices for data and software (D/S) citation? Critical Actions were identified as follows:

1. Identify the workflow from origin to citable outcomes (each person and domain needs to understand the workflow).
2. Request that publishers and repositories interlink their platforms and processes so that article references and D/S citations cross-reference each other. Make commitments to move forward on achieving this. Recognition and awareness is important.
    a. Actors: Repositories (data centers, virtual/digital, other repositories) AND publishers through their professional associations.
    b. Request funding agency support to accomplish this interlinking.

3. Request that the research community develop a primary consistent data and software citation record format (e.g., analogous to BibTex or RIS bibliography formats used in journal publishing) to support D/S citation. Journals and professional societies need to take a more active role in curating citation style files (i.e., actually curate their styles and pay attention to how Zotero or CiteProc bug tracker handle similar files).

3.2 Beyond the University Model

This session addressed the question: What are the challenges with the university model and reward structure with respect to D/S citation (and should we go beyond it)? Critical Actions were identified as follows:

1. A public/private consortium should create a software DMZ for universities (and others) to use to sustain ownership of community software as infrastructure.
2. A philanthropic foundation should raise $250 million for an endowment to sustain funding of open software and data sets based on community usage.
3. A professional society, such as the Association for Computing Machinery or the Institute of Electrical and Electronic Engineers, should lobby university provosts to recognize software in tenure processes and to provide compensation for faculty and staff reflective of market rates (perhaps citing the Software Sustainability Institute manifesto[5]).
4. The appropriate program officers at federal agencies should fund a program in which only a PI's contributions to data collections and software are considered (and papers are specifically excluded).
5. Federal agencies and universities should fund "Software Carpentry"[6] workshops and provide time for students, developers, and researchers in academia to learn and adopt appropriate software engineering skills.
6. Software engineers and data specialists should establish a "Science of Team Science" that will define essential procedures for software and data production and maintenance, as well as the groups within organizations that should be responsible for tasks within those procedures. These procedures should become criteria for funding.

3.3 Computational Reproducibility

This session addressed the question: How can we enable D/S citation for computational reproducibility? Critical Actions were identified as follows:

1. Reproducibility labeling: Adopt a public labeling system for articles that illustrates how well each article adheres to reproducibility practices. Then build a reputation system around it.
2. Reproducibility infrastructure: With partners (including industry) build a Web-based infrastructure for codes and software that stores objects and provenance and is described using community standards for provenance and workflows.
3. Reproducibility training: Dedicate funding to graduate-level training for scientists in reproducibility (computational skills and fundamental scientific method).
4. Reproducibility funding: Dedicate a portion of a research program's funding to reproducibility studies, comprised of reproduction of recently published results.

[5] See Software Sustainability Institute Manifesto, http://www.software.ac.uk/policy/manifesto.

[6] See Training on Software practices for Researchers, http://software-carpentry.org.

3.4 Difference between Software and Data

This session addressed the question: What are the differences between software and data with respect to citation? Critical Actions were identified as follows:

1. Build a conceptual (metadata) framework that supports differentiation of data and software, as well as their common elements. This will aid in understanding characteristics of both.
2. Increase understanding of use cases that illustrate the differences between software and data. Compare software and data in these use cases and establish why both (software and data) should or should not require a separate research object (RO).
3. Establish standard points-in-life for the creation of a Digital Object Identifier (DOI) or other persistent identifier that clarifies which pieces of a RO need a DOI.

3.5 Identify and Version

This session addressed the question: How can we identify authors and version contributions for D/S citation? Critical Actions were identified as follows:

1. Federal funding agencies should include D/S citation tools in their definitions of cyberinfrastructure and provide funding accordingly.
2. DCIG, ESIP, and other organizations that deal with data and software should review the reports and recommendations of the Research Data Alliance (RDA) related to persistent identifier (PID) implementations for dynamic data[7] and assess if these can be endorsed.
3. Establish a joint declaration of principles, possibly coordinated by FORCE11, first for general identifiers and then for individual research communities (e.g., ESIP, RDA). Allow each community to assess its own compliance options (e.g., PIDs such as ORCID IDs[8] for people and software).
4. Request that an organization (as yet unidentified) develop guidelines for trusted software repositories for science (similar to trusted digital data repositories).
5. The Coalition for Publishing Data in the Earth and Space Sciences (COPDESS) should add language to their Statement of Commitment and WDS/RDA Data Publishing Interest Group should agree that publishers and repositories should implement systems supporting nondeterministic DOI assignment and corresponding metadata generation (i.e., the order that these actions happen shouldn't matter).

3.6 Interoperable Frameworks

This session addressed the question: How can we create interoperable frameworks for D/S citation? Critical Actions were identified as follows:

1. Ask federal funding agencies to require every PI to have a permanent human identifier (e.g., ORCID, which resolves critical issues of identifying individuals).
2. Coordinate an agreed metadata model for both software and data so that each repository can define its profile of that metadata model.
3. At a global level, establish a "Scientific Solutions Center" (a system of systems) supported by a common (REST) API that brokers between trusted, distributed software

[7] See https://www.rd-alliance.org/groups/data-citation-wg.html.
[8] See http://orcid.org.

and data repositories to better support “Scientific Discovery through Agreed Norms and Practices for Software and Data Citation and Attribution.”
4. Focus resources on bringing together (coordinating and funding) experienced experts to enable greater interoperability and searchability across repositories of scientific data and software objects.

3.7 Large and Complex Data

This session addressed the question: How can we address special issues with large data sets and computational data? Critical Actions were identified as follows:

1. Fund repositories in order to fully:
    a. Participate in developing broad, community-sanctioned (e.g., Research Data Alliance Working Group) recommendations, including definitions and mutability models for content, versioning, identifier assignments, and related aspects.
    b. Implement community sanctioned recommendations, for example, recommendations on assigning PIDs to dynamic queries.[9]
2. Require identifier registration authorities to participate in science and big dynamic data community efforts, rather than simply addressing the issue from the library perspective.

3.8 Social Barriers

This session addressed the question: What are some strategies for breaking down the social barriers associated with D/S citation? The discussion focused on breaking down social barriers while also developing a better understanding of the culture of sharing among researchers and how can this be leveraged more effectively. Critical Actions were identified as follows:

1. Conduct outreach on, and advocate for, the value of software as part of the scientific method. These efforts should be aimed at:
    a. Early career researchers: Provide training on D/S citation and practices as part of a curriculum. University buy-in is necessary for this effort to succeed.
    b. Global academic societies, journals, reviewers, and related communities.
    c. University administrators, who need to better track their intellectual contributions and assets in terms of data and software.

    Templates and best practices should be provided to all communities as part of this outreach.
2. Separate sharing and citation when looking for solutions and providing guidelines (separate the organization of data and software contributions from the means of organizing them).
3. Provide funding for systems’ interoperability, and provide tools and templates to facilitate D/S citation.

[9] See https://www.rd-alliance.org/groups/data-citation-wg/wiki/scalable-dynamic-data-citation-rda-wg-dc-position-paper.html.

3.9 Subsetted, Derived, and Aggregated

This session addressed the question: How should we cite subsetted, derived, and aggregated data and software? Critical Actions were identified as follows:

1. Data and software repository landing pages should describe the full provenance of the data using appropriate standards (e.g., DataCite, W3C PROV, DC, or W3C DCAT).
2. Authors should be able to cite data and software in their articles at an appropriate level of granularity.
3. To validate and achieve points 1 and 2, a pilot program with selected repositories should be funded.
4. Request that the DataCite Metadata Working Group enrich metadata schema to support description of collections.
5. Launch a mini project showing how repositories can enable authors to correctly cite data (including data generated on-the-fly) subsets and aggregations using existing repositories and services.
6. Fund a follow-up activity to develop tools that allow traversing up and down citation chains.

3.10 Useful Metrics

This session addressed the question: How can we establish useful D/S metrics? Critical Actions were identified as follows:

1. Federal funding agencies should support an effort to convene key players to identify and harmonize standards on roles, attribution, value, and transitive credit (in an extensible framework). All key sponsors would be recognized.
2. Agencies, publishers, societies, and foundations should fund implementation grants to identify and measure data and software impacts in a way that is relevant to stakeholders and research communities.
3. Identify a model to iterate and improve the standards framework.
4. Define discovery and use metadata standards for software.
5. Break down “contributorship” to become more nuanced to go beyond traditional authorship and become something more akin to film credits.

4. Impact and Likelihood of Implementing Critical Actions

Following the breakout sessions, workshop participants reconvened and distributed themselves among eight tables. A representative from each breakout group presented the proposed Critical Actions for their topic. After each presentation, participants at each table were given several minutes to reach consensus on where each Critical Action fits in a “How, Wow, Now, Why?” matrix, a technique used to gather feedback about the anticipated impact and likelihood of each Critical Action. Figure 2 shows how these categories were aligned, with “Likelihood” on the horizontal axis and “Impact” on the vertical axis.

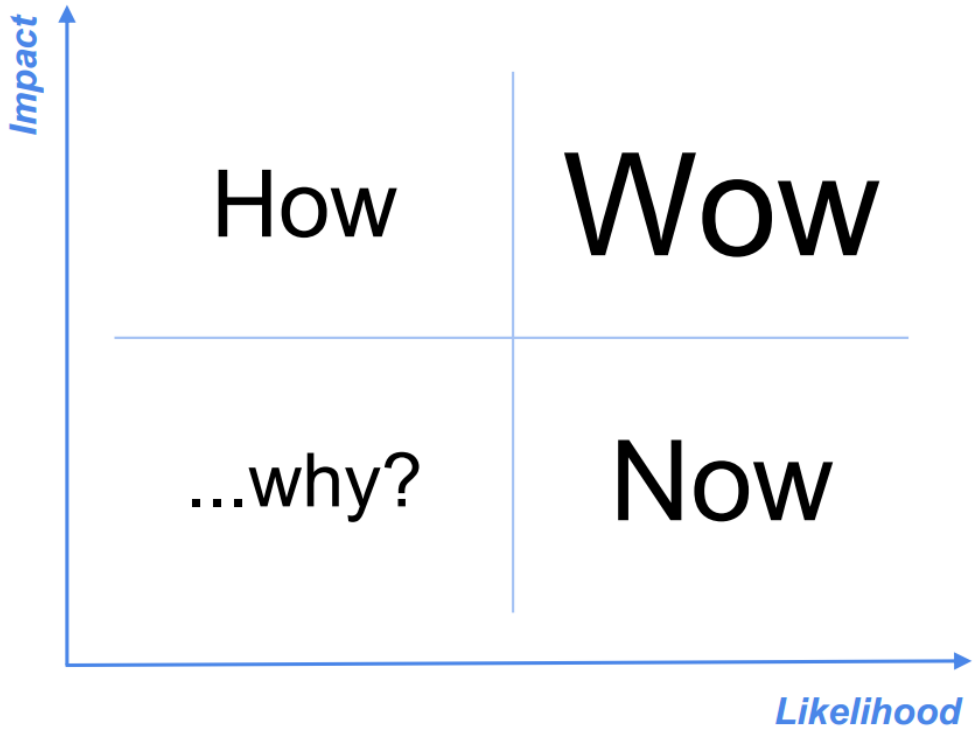

Figure 2: The "How, Wow, Now, Why?" matrix used to gather participant feedback on the impact and likelihood of each proposed Critical Action.

The terms in the matrix can be summarized as follows:

- *How:* Critical Action has a moderate likelihood of actionable realization, but high potential impact.
- *Wow:* Critical Action has a high likelihood of actionable realization and high potential impact.
- *Now:* Critical Action has a high likelihood of actionable realization, but moderate impact.
- *Why?:* Critical Action has a low likelihood of actionable realization and low perceived impact.

The workshop facilitators used a poster-size version of the How, Wow, Now, Why? matrix and attached post-it notes (with each Critical Action written on a note) to visually convey participants' views about the impact and likelihood of each Critical Action. The results of this process are shown in Figure 3, where each post-it note represents a Critical Action and different colors represent different breakout topics. Table 1 shows how each participant table voted regarding the position of each Critical Action in the matrix.

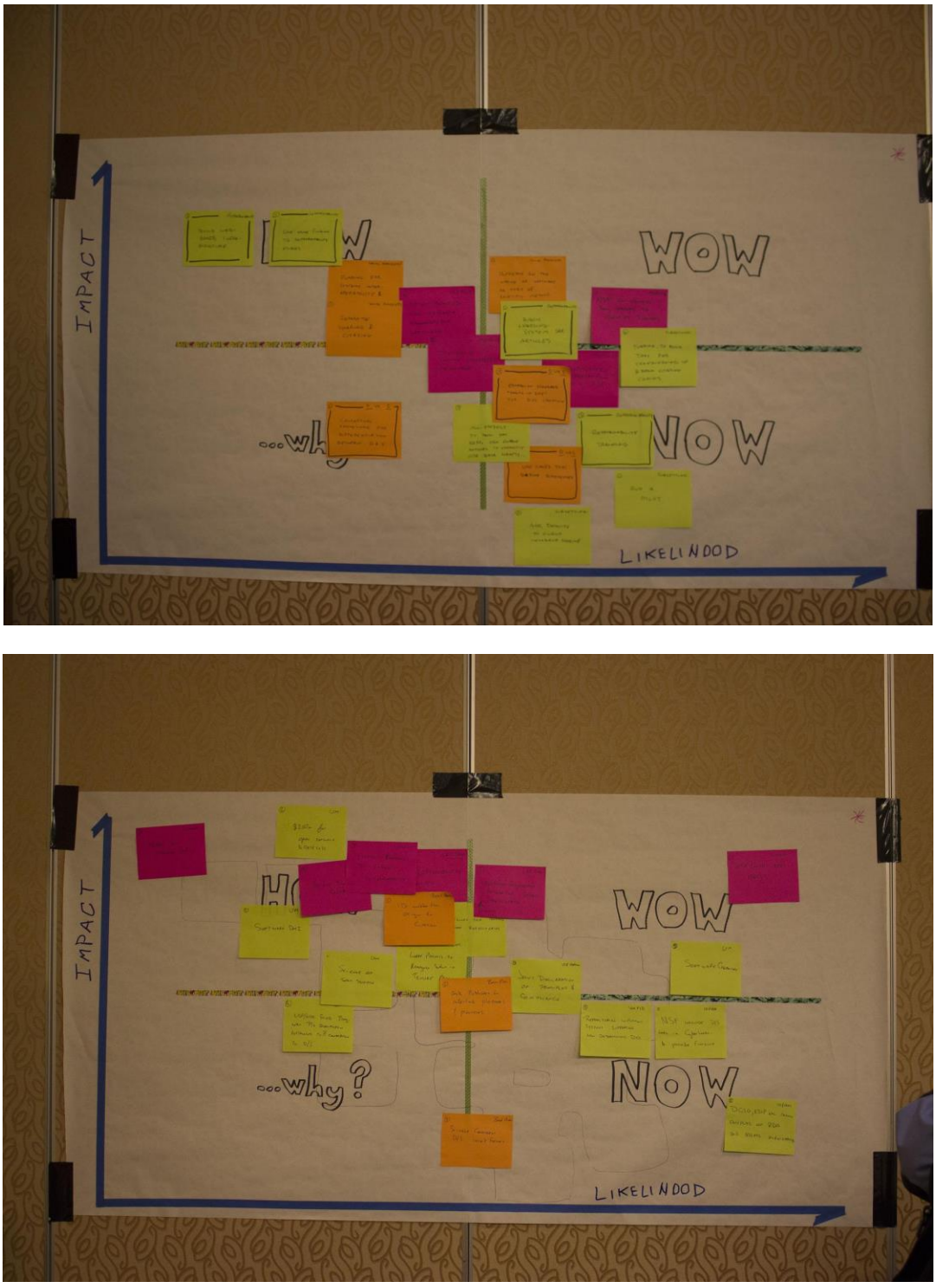

Figure 3: Posters showing how participants sorted the proposed Critical Actions based on the "How, Wow, Now, Why?" matrix for breakout sessions on January 29 (top) and 30 (bottom).

Table 1: The results of the “How, Wow, Now, Why?” sorting exercise.

Use-case Topic and Critical Asks	HOW	WOW	NOW	WHY?	Groups that reached HOW, WOW, NOW, or WHY? consensus (max: 8)
Best Practices					
Identify the workflow from origin to citable outcomes	4	2	2		8
Ask publishers and repositories to interlink their platforms or processes	3	2	3		8
Single consistent data and software citation record format		1	4	3	8
Beyond the University Model					
Public/private consortium should create a software dmz for universities (and others) to use	4	4			8
Raise 250 million dollars for an endowment for sustainability funding of open software and datasets	6	2			8
Lobby Provosts to recognize software in tenure process	4	1	3		8
NSF/NIH fund a program where the PI’s qualifications are restricted to the PI’s contributions to data collections and software	2	1		5	8
Feds/Universitiies fund Software Carpentry and time		5	2	1	8
Science fo Team Science essential characteristics of appropriate orgs for software/data maint	4		1	2	7
Computational Reproducibility					
Reproducibility labeling	3	3	2		8
Reproducibility infrastructure	6		2		8
Reproducibility training	1	2	4		7
Reproducibility funding	4	1		2	7
Differences between Data vs. Software					
Conceptual (metadata) framework that supports differentiation of data and software	1	2	1	4	8
Use cases that drive the differences			6	2	8
Establish standard points-in-life for “DOI” creation what pieces of a RO needs a DOI	3	1	4		8
Identify & Version					
NSF should must include D&S citation tools in its definition of CyberInfrastructure and provide funding accordingly		2	5	1	8
DCIG, ESIP, etc. need to review outputs of RDA related to PID implementations and assess if these can be endorsed			7	1	8
Propose joint declaration of principles (FORCE11 could coordinate) for general identifiers and then individual communities (e.g., ESIP, RDA) can assess compliant options for their community (e.g. IDs for people, SW)	2	3	3		8
Ask an organization (not sure which?) to develop guidelines for trusted software repositories for science (similar to trusted digital data repositories)	3	2	2	1	8
COPDESS to add language to their Statement of Commitment and WDS/RDA Data Publishing Interest Group agree that repositories implement systems supporting non-deterministic DOI assignment and corresponding metadata generation (i.e. order that these actions happen shouldn’t matter)		1	5	1	7
Interoperable Frameworks					
NSF & NIH to require every PI to have an human identifier	1	6	1		8
Coordinate a model for a Metadata Profile so each repository	5	1	1	1	8
Scientific Solutions Center	5	2	1		8
Focused resource(s) to bring together (coordinate & fund) the expertise and experience to enable greater interoperability	5	2	1		8
Large & Complex Data					
Fund repositories to fully participate in development of broad community-sanctioned (e.g RDA WG) recommendations including definitions and mutability models for content, versioning, and identifier assignments, etc.; and implement community sanctioned recommendations	4	3	1		8
Identifier registration authorities should participate in science and big dynamic data	2	4	1	1	8
Social Barriers					
Outreach on the value of software	3	2	3		8
Separate sharing and citation when looking for solutions	5	1		2	8
Funding for systems interoperability	4	1	1	2	8
Subsetted, Derived & Aggregated					
Pilot with selected repositories	1	2	5		8
Ask DataCite metadata working group to enrich metadata schema			5	2	7
Mini project showing how repositories can enable authors to correctly cite data	1		4	1	6
Funding for tool that allows traversing up and down citation chains	1		4	2	7
Useful Metrics					
Convene players for standards	1	2	4	1	8
Fund implementation grants relevant to stakeholders	2	1	4	1	8
Identify model and improve citation standard framework	2	1	1	2	6
Define discovery/use metadata standards for software	4		4		8

5. Summary and Next Steps

The workshop generated substantial interest and excitement among participants. A majority of the workshop participants agreed that now is the time to move beyond workshops and discussions on D/S citation and begin implementing actions articulated at the workshop. If implemented, these Critical Actions can guide further progress in data and software citation and attribution. Participants expressed great interest in advancing pilot programs to help research communities implement practices and procedures that facilitate improved credit, measurement, and attribution of research. As a next step, NSF attendees encouraged groups to submit proposals that elaborate on these ideas.

Acknowledgements

This workshop was funded by the National Science Foundation through grant number 1448360 with additional support from the Alfred P. Sloan Foundation. The organizing committee expresses its sincere appreciation to the National Science Foundation and Sloan Foundation for their support. The committee extends a special thank you to all the attendees for participating in the workshop discussions and panel sessions, leading breakout sessions, taking notes, and contributing to this report. The interest in pursuing a coordinated effort among research communities and related projects is now stronger than ever as a result of this workshop. The consensus built, insights gained, and recommendations provided will be of great value to the NSF, our respective research communities, industry partners, and other government agencies.

Appendix A: Workshop Participants

Stan	Ahalt	RENCI, University of North Carolina at Chapel Hill; NCDS
Phatty	Arbuckle	Massachusetts Institute of Technology
Karl	Benedict	University of New Mexico
Vivien	Bonazzi	National Institutes of Health
Abigail	Cabunoc	Mozilla Foundation
Bonnie	Carroll	Information International Associates
Tom	Carsey	Odum Institute, University of North Carolina at Chapel Hill
Cyndy	Chandler	Woods Hole Oceanographic Institution
Ishwar	Chandramouliswaran	National Institutes of Health
Sayeed	Choudry	John Hopkins University
Tim	Clark	Harvard University
Alva	Couch	CUAHSI; Tufts University
Jon	Crabtree	Odum Institute, University of North Carolina at Chapel Hill
Mercè	Crosas	Harvard University
Ruth	Duerr	NSIDC, University of Colorado at Boulder
Ian	Fore	National Institutes of Health
Victoria	Forlini	American Geophysical Union
Harry	Furukawa	American Geophysical Union
Yolanda	Gil	ISI, University of Southern California
Jane	Greenberg	Drexel University
Josh	Greenberg	Sloan Foundation
Paul	Groth	Elsevier
Sophie	Hou	University of Illinois at Urbana-Champaign
James	Howison	University of Texas at Austin
Leslie	Hsu	Columbia University
Lorraine	Hwang	University of California at Davis
Luis	Ibanez	Google, Inc.
Ray	Idaszak	RENCI, University of North Carolina at Chapel Hill
Matthew B.	Jones	NCEAS, University of California at Santa Barbara
Dan	Katz	National Science Foundation
Jens	Klump	Commonwealth Scientific and Industrial Research Organization
Madison	Langseth	University of Tennessee at Knoxville
Jennifer	Lin	Public Library of Science
Maryann	Martone	University of California at San Diego
Eric	Moran	SAGE Publishing
Fiona	Murphy	John Wiley & Sons, Inc.
Mark	Parsons	Research Data Alliance; Rensselaer Polytechnic Institute
Erin	Robinson	Foundation for Earth Science
Sweitze	Roffel	Elsevier
Mark	Schildhauer	NCEAS, University of California at Santa Barbara
Kes	Schroer	Dartmouth University
Sudhir	Shrestha	National Oceanic and Atmospheric Administration
Joan	Starr	CDL; University of California Office of the President
David	Tarboton	Utah State University
Mike	Taylor	Elsevier
Curt	Tilmes	National Aeronautics and Space Administration
Nic	Weber	University of Illinois at Urbana-Champaign
Robert	Wolfe	National Aeronautics and Space Administration
Lesley	Wyborn	National Computational Infrastructure; Australian National University