Digital Library Committee Project:  
Best Practices Guidelines for   
Data, Software, and Reproducibility in Publication

Draft: 4 Feb 2016

# Goals and Scope

The goal of this project is to design best practices for open and widely acceptable guidelines for associating data and software artifacts with published articles in collaboration with other publishers and academics. Reviewing and providing access to these software and data artifacts should improve reliability and reproducibility of the archival computing literature.

These guidelines do not anticipate that ACM will provide storage or repository services for artifacts unless those artifacts are specifically associated with published articles. Other services, such as [github](https://github.com/) and [IPCSR,](https://www.icpsr.umich.edu/icpsrweb/landing.jsp) are better equipped to provide such repository services than the ACM Digital Library, with its mission of making archival computing research discoverable and accessible while fully integrating bibliographic data of all computing literature.[[1]](#footnote-0)

## Background

Concerns about reproducibility of research have been covered widely in the popular press recently. Emilie Marcus, from Cell Press, notes that this emphasis on reproducibility is not new and that it parallels budgetary constraints on research. She notes that in the budget-tightened 1980s, the US federal government emphasized training on experimental methods in its grants, and created the Office of Research Integrity to address perceived problems in reproducibility. [[2]](#footnote-1)

The inability to reproduce research results is an issue of concern in many scientific fields. For example, in oncology research one study failed to replicate the results of 47 out of 53 major papers.[[3]](#footnote-2) In psychology, The Reproducibility Project was recently undertaken to replicate the results of 97 studies published in highly regarded journals, in which only 36% were significantly replicable and that even in those studies that were replicable, the effect of the interventions were about half of what the studies claimed.[[4]](#footnote-3)

An economics paper that argued that countries whose national debt exceeded 90% of GDP had experience low economic growth was cited by policy experts and politicians to justify spending cuts and tax increases for two years before it was discovered to be in error by researchers using the same data.[[5]](#footnote-4) In fact, a recent study succeeded in replicating results from only 29 out of 67 economics papers even when using data and code provided by the original authors. Some, like Daniel Shanahan, have called for a radical transformation of the journal article, and deemphasizing results over process. One result could be increased replicability, although questions remain about how to associate all of the publications that correspond to the same research.[[6]](#footnote-5)

Publishers in many disciplines are turning to the problem of reproducibility, as evidenced by a recent International Scientific Technical and Medical Publishers Association (STM) innovation seminar on the topic.[[7]](#footnote-6)

This document addresses the following areas that will be necessary to support the inclusion of data and software artifacts with publications:

* description and metadata
* review process
* author and reviewer motivation
* storage and long-term preservation
* citation and identification
* document structure

This document also includes a section on problems of associating artifacts with publication and suggests some possible solutions and areas for further exploration.

In 2013, the National Information Standards Organization (NISO) released guidelines for supplemental materials for journal articles.[[8]](#footnote-7) The recommended practices were intended to apply across all disciplines. The report notes, “...[M]embers, have been cognizant of the differing cultures and practices across scholarly publishing and aware of the fact that different fields have different requirements. Consequently, the recommendations include many gray areas.” Similarly, during the process of developing these ACM guidelines, it became clear that different computing communities today have different approaches towards the problems and steps to improving the reliability and reproducibility of the computing literature. Representatives from a broad array of these communities contributed to these guidelines. [Appendix A](#h.v5stfv1kjc7w) provides descriptions of approaches to these problems in several communities. In addition, [Appendix B](#h.11ho6yw93vjd) includes a list of the participants that participated in the development of these guidelines.

Data and its open availability is becoming increasingly important in the UK. In part this is because the government has started to make certain demands as to data availability as a follow on from open access publication. All Research Council UK (RCUK) funders insist on a data plan including long-term storage and persistent endpoints before a grant submission can occur. In this case I thought it might be useful to add the salient items of the University of Manchester requirements for later discussion. [http://www.library.manchester.ac.uk/services-and-support/staff/research/ser…](http://www.library.manchester.ac.uk/services-and-support/staff/research/services/research-data-management/policy/) [http://www.library.manchester.ac.uk/services-and-support/staff/research/ser…](http://www.library.manchester.ac.uk/services-and-support/staff/research/services/research-data-management/data-management-planning/)

The RCUK has set out a set of general guidelines for data management, which are summarized below:

* Publicly funded research data should be made openly available.
* Data management policies and plans should use standards and best practices. Data with long-term value should be preserved.
* Public metadata should enable other researchers to understand the research and reuse potential of the data. Published results should include data access information.
* Legal, ethical and commercial constraints should be considered.
* Limted time data embargoes are allowed.
* Data sources should be correctly attributed and data should be used in conjunction with any licenses or restrictions.
* Public funds can support management and sharing of publicly-funded research data[[9]](#footnote-8)

Institutions managing RCUK funded projects create more specific research data management policies and data management plans. Examples can be found at the University of Manchester.[[10]](#footnote-9),[[11]](#footnote-10)

# Definitions

Different sub-communities within ACM and in the academic community as a whole have subtle differences in meaning for the same term. In this document, we recommend the following definitions, drawn, when possible, from the cited standards.

## Artifact

In this document, the term *artifact* is used to refer to software, data, and other supplemental material to a publishable article that is additional or related to, but not integral to understanding the article. NISO draws a distinction between *integral, additional,* and *related* content. Its guidelines exclude the third category, which contains material not essential to understand the article in question but that could help researchers reproduce the results.[[12]](#footnote-11) By contrast, these guidelines consider related content to be essential to the purpose of allowing results to be reproduced by third parties.

## Openness

“Open data is data that meets the criteria of intelligent openness. Data must be accessible, useable, assessable and intelligible.”[[13]](#footnote-12)

## Replication

Computational biologist Roger Peng writes,“Replication is the ultimate standard by which scientific claims are judged. With replication, independent investigators address a scientific hypothesis and build up evidence for or against it.”[[14]](#footnote-13) Yet, as datasets become larger and more complex, he continues, it is increasingly difficult to meet that standard because of the time and expense of recreating the research. A lower standard of reproducibility, he argues, is a minimum standard where the software and data are made available to third parties so that they can verify the original results. In Peng’s words, “The standard of reproducibility calls for the data and the computer code used to analyze the data be made available to others. This standard falls short of full replication because the same data are analyzed again, rather than analyzing independently collected data.”[[15]](#footnote-14)

In computational mathematics, on the other hand, replicating the experiment means “making available to others the associated data and code. In this field, there is not as clear a distinction between replicability and reproducibility. “The ‘reproducible research’ movement recognizes that traditional scientific research and publication practices now fall short of this ideal [of replicability], and encourages all those involved in the production of computational science – scientists who use computational methods and the institutions that employ them, journals and dissemination mechanisms, and funding agencies – to facilitate and practice really reproducible research.”[[16]](#footnote-15)

The taxonomy to the Institute for Computational and Experimental Research in Mathematics (ICERM) report on Reproducibility draws the distinction of replicability and reproducibility based on who has access to the data and code. Replicable research makes the data and code available just to the reviewers, whereas reproducible research (also called “open research”) makes it openly available to the public.

### ISO 3534-2:2006

“replicate sampling  
<bulk material> [experimental sampling (5.2.3)](https://www.iso.org/obp/ui/#iso:std:iso:3534:-2:ed-2:v1:en:term:5.2.3) where [increments (5.2.7)](https://www.iso.org/obp/ui/#iso:std:iso:3534:-2:ed-2:v1:en:term:5.2.7) are taken simultaneously or consecutively in pairs, or in higher order multiples in order to constitute multiple [composite samples (5.3.4)”](https://www.iso.org/obp/ui/#iso:std:iso:3534:-2:ed-2:v1:en:term:5.3.4)[[17]](#footnote-16)

## Repeatability

This document uses *repeatability* to mean that the same results are obtained in the same facility, by the same operator, using the same data and equipment within a short period of time as the original experiment or procedure.

### [ISO 3534-2:2006

“3.3.6 repeatability conditions

observation conditions where [independent test/measurement results (3.4.3)](https://www.iso.org/obp/ui/#iso:std:iso:3534:-2:ed-2:v1:en:term:3.4.3) are obtained with the same method on [identical test/measurement items (1.2.34)](https://www.iso.org/obp/ui/#iso:std:iso:3534:-2:ed-2:v1:en:term:1.2.34) in the same test or measuring facility by the same operator using the same equipment within short intervals of time

Note 1 to entry:

Repeatability conditions include:

* — the same measurement procedure or test procedure;
* — the same operator;
* — the same measuring or test equipment used under the same conditions;
* — the same location;
* — repetition over a short period of time.”[[18]](#footnote-17)

### JCGM Working Group 2

“2.20 (3.6, Notes 1 and 2)

repeatability condition of measurement

repeatability condition

condition of measurement, out of a set of conditions that includes the same measurement procedure, same operators, same measuring system, same operating conditions and same location, and replicate measurements on the same or similar objects over a short period of time

NOTE 1 A condition of measurement is a repeatability condition only with respect to a specified set of repeatability conditions.

NOTE 2 In chemistry, the term “intra-serial precision condition of measurement” is sometimes used to designate this concept”[[19]](#footnote-18)

## Reproducibility

The ICERM Reproducibility Report uses the terms “reproducible research” and “reproducibility” most often refer to the ability to recreate computational results from the data and code used by the original researcher.[[20]](#footnote-19) In this document, we use the term reproducibility to mean that the same results are achieved using the same method and data but in a different facility using different equipment and conducted by different operators than the original experiment or procedure.

### [ISO 3534-2:2006

“3.3.10

reproducibility

[precision (3.3.4)](https://www.iso.org/obp/ui/#iso:std:iso:3534:-2:ed-2:v1:en:term:3.3.4) under [reproducibility conditions (3.3.11)](https://www.iso.org/obp/ui/#iso:std:iso:3534:-2:ed-2:v1:en:term:3.3.11)

Note 1 to entry: Reproducibility can be expressed quantitatively in terms of the dispersion [characteristics (1.1.1)](https://www.iso.org/obp/ui/#iso:std:iso:3534:-2:ed-2:v1:en:term:1.1.1) of the results.

Note 2 to entry: Results are usually understood to be corrected results.

3.3.11

reproducibility conditions

observation conditions where [independent test/measurement results (3.4.3)](https://www.iso.org/obp/ui/#iso:std:iso:3534:-2:ed-2:v1:en:term:3.4.3) are obtained with the same method on [identical test/measurement items (1.2.34)](https://www.iso.org/obp/ui/#iso:std:iso:3534:-2:ed-2:v1:en:term:1.2.34) in different test or measurement facilities with different operators using different equipment”[[21]](#footnote-20)

### JCGM Working Group 2

“2.24 (3.7, Note 2)

reproducibility condition of measurement

reproducibility condition condition of measurement, out of a set of conditions that includes different locations, operators, measuring systems, and replicate measurements on the same or similar objects

NOTE 1 The different measuring systems may use different measurement procedures.

NOTE 2 A specification should give the conditions changed and unchanged, to the extent practical.”[[22]](#footnote-21)]

Emilie Marcus defined reproducibility as the condition where “other scientists using the same materials and conditions can reproduce the work” whereas she calls the generalizability of reproducible results “robust.”[[23]](#footnote-22)

# 3. Stakeholder Motivation

According to Zoltain Dienes, “Science has an incentive problem.” In other words, what is good for science, namely doing good quality research regardless of the outcomes, is not necessarily good for scientists, who cannot publish unless they find positive outcomes.[[24]](#footnote-23)Creating a cultural environment and process of sharing data and software requires additional effort on the part of already busy editorial boards, reviewers, and authors. This section describes best practices that can be used to motivate each of these stakeholder groups to provide the resources necessary to implement them.

The first recommendation of the ICERM report was to promote a cultural change supporting reproducibility, and the second was that journals, funding agencies, and employers should support this cultural change.[[25]](#footnote-24)

“Statements from societies and others on the importance of reproducibility could advance the culture change.”[[26]](#footnote-25)

Also see this article from *Science.[[27]](#footnote-26)*

Science Code Manifesto is an attempt to encourage researchers to make their code available publically by endorsing its principles. In particular the code principle states, “All source code written specifically to process data for a published paper must be available to the reviewers and readers of the paper.”[[28]](#footnote-27)

## Authors

The Engineering and Physical Sciences Research Council (EPSRC) of the UK has posted clarification of its expectations for research data management.[[29]](#footnote-28)

In Expectation I, institutions are instructed to explain the importance and legal requirements of sharing research data to their researchers.

EPSRC Expectation II. “Published research papers should include a short statement describing how and on what terms any supporting research data may be accessed.”

“Researchers should be encouraged to recognize the potential benefits of openness and reproducibility,” ICERM.

“Appropriate credit should given for code and data contributions including an expectation of citation.”

“Another suggestion is to instantiate yearly award from journals and/or professional societies, to be awarded to investigators for excellent reproducible practice”[[30]](#footnote-29)

SIGMOD outlines these benefits to authors to participate in their Reproducibility project.

1. Easing the way for other researchers to compare an author’s work, and adopting and extending that work creates greater recognition and higher impact. 
2. Participating in the SIGMOD Reproducibility enables authors’ papers to display a “Reproducible” label in the ACM digital library.
3. Reproducible papers will also be advertised at DBworld.
4. A separate SIGMOD Reproducibility website hosts and publicizes Reproducible papers in a centralized location where researchers will be able to find all the experimentation material of sharable SIGMOD papers.
5. Reinforces good research practices so that compliance should not require additional effort.[[31]](#footnote-30) “Repeatable experiments simplify your own work (and help others to understand it better).”[[32]](#footnote-31)

Data-only journals like *Genomics* and *Data in Brief* from Elsevier give authors publication credit for sharing research data.[[33]](#footnote-32) Another list of data-only journals can be found at the TRAC website.[[34]](#footnote-33)

Several research approaches to reduce the bias for finding significant results suggested by Dienes include the following. He suggests that these types of studies could be encouraged by journals.

1. Anonymize data sets and make the analyst “blind” to which is the real data set.
2. Create different teams to run experiments on the same data with different approaches
3. Introduce adversarial collaboration, where proponents with opposite views get together to agree on a protocol and publish together.
4. Conduct and publish high-value direct replication studies.[[35]](#footnote-34)

## Reviewers

Publishers to provide a “set of standards for reviewing papers” including best practices and a rational procedure for allowing exceptions or exclusions.

Referees need access to computer code or data, and the ability to run computational tests of the results in submitted papers, if desired.

The computational claims of a manuscript to be verifiable at another site such as RunMyCode.org, or on another computer system with a similar configuration.

Issues to resolve:

1. anonymous versus public review,
2. persistence (longevity) of code and data that is made publicly available,
3. how code and data can be “watermarked,” so that instances of uncited usage (plagiarism) can be detected and provenance better established,
4. how to adjudicate disagreements that inevitably will arise.

Higher rigor would be required of papers making “important assertions, such as the computer-assisted proof of a long-standing mathematical result, new scientific breakthroughs, or studies that will be the basis for critical policy decisions.”[[36]](#footnote-35)

Providing reviewers a script to install the environment makes this task less onerous.[[37]](#footnote-36)

Reviewers could reject the papers unless the data is online and publicly available. The *American Journal of Political Science* has begun to use public data to try to reproduce the results as part of the review process.[[38]](#footnote-37)

## Institutions

See EPSRC Expectation I in the [Authors](#h.60vls9cdsp4a) section above.

EPSRC “Expectation III

Each research organisation will have specific policies and associated processes to maintain effective internal awareness of their publicly‐funded research data holdings and of requests by third parties to access such data; all of their researchers or research students funded by EPSRC will be required to comply with research organisation policies in this area or, in exceptional circumstances, to provide justification of why this is not possible.”

EPSRC “Expectation IV

Publicly‐funded research data that is not generated in digital format will be stored in a manner to facilitate it being shared in the event of a valid request for access to the data being received (this expectation could be satisfied by implementing a policy to convert and store such data in digital format in a timely manner);”

“Expectation VIII

Research organisations will ensure that effective data curation is provided throughout the full data lifecycle, with ‘data curation’ and ‘data lifecycle’ being as defined by the Digital Curation Centre. The full range of responsibilities associated with data curation over the data lifecycle will be clearly allocated within the research organisation, and where research data is subject to restricted access the research organisation will implement and manage appropriate security controls; research organisations will particularly ensure that the quality assurance of their data curation processes is a specifically assigned responsibility.”

“Expectation IX

Research organisations will ensure adequate resources are provided to support the curation of publicly‐funded research data; these resources will be allocated from within their existing public funding streams, whether received from Research Councils as direct or indirect support for specific projects or from higher education Funding Councils as block grants.”

From ICERM:

1. Tenure
   1. Software and dataset contributions should be rewarded as part of expected research practices.
   2. Data and code citation practices should be recognized and expected in computational research.
   3. Prizes for reproducible research should be recognized in tenure and promotion decisions.
2. Libraries
   1. Institutional libraries should provide template data management plans
   2. Institutional Libraries should provide archiving, stewardship and dissemination of scholarly objects.
   3. Increase coordination between departments and the institute’s library system.[[39]](#footnote-38)
3. Teaching
   1. Formats
      1. full academic courses,
      2. incorporation into existing courses,
      3. workshops and summer schools,
      4. online courses or self-study materials,
      5. teaching-by-example on the part of mentors.[[40]](#footnote-39)
   2. Topics
      1. “version control and use of online repositories,
      2. modern programming practice including unit testing and regression testing,
      3. maintaining “notebooks” or “research compendia”,
      4. recording the provenance of final results relative to code and/or data,
      5. numerical / floating point reproducibility and nondeterminism,
      6. reproducibility on parallel systems,
      7. dealing with large datasets,
      8. dealing with complicated software stacks and use of virtual machines,
      9. documentation and literate programming,
      10. IP and licensing issues, proper citation and attribution”[[41]](#footnote-40)
4. Encourage and teach the use of tools to aid in reproducibility.[[42]](#footnote-41)

## Funders

Government and private funding agencies and private foundations should establish reasonable standards for proposals to encourage reproducibility.

1. Software and data should be “open by default” unless it conflicts with other considerations.

2. Proposals involving computational work might be required to provide details such as

• Extent of computational work to be performed.

• Platforms and software to be utilized.

• Reasonable standards for dataset and software documentation, including reuse (some agencies already have such requirements [8]).

• Reasonable standards for persistence of resulting software and dataset preservation and archiving.

• Reasonable standards for sharing resulting software among reviewers and other researchers.

3. Add “reproducible research” to the list of specific examples that proposals include in their requirements such as “Broader Impact” statements.

4. Software and dataset curation should be explicitly included in grant proposals and recognized as a scientific contribution by funding agencies.

5. Make available templates for data management plans.

6. Provide support for training workshops on reproducibility.

7. Provide support for cyberinfrastructure for reproducibility at scale, for both large projects and long-tail research efforts. [[43]](#footnote-42)

## Publishers

One approach taken by publishers has been to create “Registered Reports” in which researchers submit their hypotheses, experimental design, and procedures to peer review *before* collecting any data.[[44]](#footnote-43) Publishers may also sign onto and publicize efforts like the #AllTrials campaign from Sense About Science, which encourages researchers to openly release data for all clinical trials through a variety of mechanisms now becoming available. They may also adhere to the International Council of Medical Journal Editors (ICMJE) policy on registration, and commit to adding registration metadata (such as Registration ID, and dates for registration, trial beginning and end, registry submission and article submission.) Journals can highlight trials where data has not been registered, encourage authors to openly share data, and audit and publish the results of these policies.[[45]](#footnote-44)

See EPSRC Expectation II in the [Author](#h.60vls9cdsp4a) section above.

See ICERM recommendation on reproducibility awards in [Author](#h.60vls9cdsp4a) section above.

Journal best practices should be created.

***Disclosure***

*As part of review process*

*Post-Acceptance*

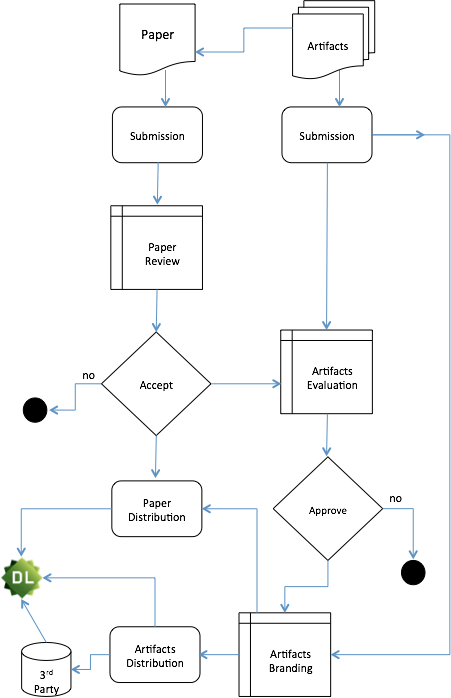
This should include “full disclosure regarding software and data use:

* specification of the dataset used (including URL and version),
* details of the algorithms employed (or references),
* the hardware and software environment,
* the testing performed, etc.,
* availability of the relevant computer code with a reasonable level of documentation and instructions for repeating the computations performed to obtain the results in the paper”[[46]](#footnote-45)

Promote standards for citing software and data within the references of a paper and not just in text.[[47]](#footnote-46)

* “A precise statement of assertions to be made in the paper.
* A statement of the computational approach, and why it constitutes a rigorous test of the hypothesized assertions.
* Complete statements of, or references to, every algorithm employed.
* Salient details of auxiliary software (both research and commercial software) used in the computation.
* Salient details of the test environment, including hardware, system software and the number of processors utilized. 9 Indeed, one needs to know which precise functions were called, with what parameter values and environmental settings? 13
* Salient details of data reduction and statistical analysis methods.
* Discussion of the adequacy of parameters such as precision level and grid resolution.
* Full statement (or at least a valid summary) of experimental results.
* Verification and validation tests performed by the author(s).
* Availability of computer code, input data and output data, with some reasonable level of documentation.
* Curation: where are code and data available? With what expected persistence and longevity? Is there a site for site for future updates, e.g. a version control repository of the code base?
* Instructions for repeating computational experiments described in the paper.
* Terms of use and licensing. Ideally code and data “default to open”, i.e. a permissive re-use license, if nothing opposes it.
* Avenues of exploration examined throughout development, including information about negative findings.
* Proper citation of all code and data used, including that generated by the authors.
* Journals or other publications could offer certifications of reproducibility that would kite-mark a paper satisfying certain requirements, as done by the journal Biostatistics, for example.
* Certification could also come from an independent entity such as RunMyCode.org.
* Journals could also create reproducible overlay issues for journals that collect together reproducible papers.
* Linking publications to sites where code and data are hosted will help shift toward reproducible research. For example, the *SIAM Journal on Imaging Science* provides cross-referencing with the peer-reviewed *journal Image Processing On Line* (IPOL) and encourage authors to submit software to IPOL. Other sites such as RunMyCode.org or Wakari might be used in a similar way.
* Finally, all code and data should be labeled with author information.”[[48]](#footnote-47)

# ACM Digital Library Workflow



## **Cases**:

1. All Artifacts and/or associated Papers are “Branded”. Distribution from DL or linked.
2. Artifacts submitted with papers:
   1. Evaluation and approval of Artifacts required for publication of paper.
   2. Artifacts are available for distribution.
   3. Artifacts are not available for distribution.
   4. Evaluation of Artifacts independent of publication of paper.
   5. Artifacts are available for distribution.
   6. Artifacts are not available for distribution.
   7. No Evaluation of Artifacts performed.
   8. Artifacts are available for distribution.
3. Artifacts only:
   1. Evaluation and approval required for distribution.
   2. No Evaluation performed.

# Data

As nicely stated in a briefing paper on journal research data at the Research Data Alliance Sixth Plenary, “In order to comply with funders, researchers need access to clear guidelines on the academic journal’s expectations when it comes to deposit and access of the supporting data. Publishers are increasingly putting in place policies to support access to data.”[[49]](#footnote-48)

Some computing communities (like Multimedia) have a culture of using standard images/videos to test algorithms. In other cases, the data to be shared so as to enable reproducibility and replicability will be unique to the paper. In either case, existing repositories may be available to store the data.

Cornell University’s Best Practices for Data Management site focuses on data and could be very helpful, particularly in the data description standards and data format standards it recommends.[[50]](#footnote-49)

Using the questions posed in the workshop agenda as a guide, the Data breakout group covered a broad set of issues relating to reproducibility of data. The discussions covered topics such as defining what constitutes data, how to cite data, and how to find data sets in digital libraries. Since the discussions covered a broad set of issues, this report first captures a summary of discussions based on the initial set of questions posed, and then presents a summary of recommendations.

**Summary of discussions based on questions posed in workshop agenda**

***Key:***

● Questions posed in the workshop agenda are in black

■ *Our discussion summaries are in blue*

## Metadata description

What metadata description standards do we need to adopt or invent for data sets?

● What Metadata description standards do we need to adopt or invent for data sets?

o For Micro data sets?

o For Macro data sets?

● What are recommended formats for submitted data?

● Should data be treated as a first-class object with its own identifier? Or as a supplemental file? If a supplement, should it have its own independent DOI and be independently citable?

■ *We should recognize at least three kinds of data and label them appropriately:*

*a.* *data sufficient to generate the plots (plot points);*

*b.* *measured data (measurements of the real world, experiments, or of a system described in the paper)*

*c.* *data inputs necessary to run simulations*

■ *The general idea is to have one object with Metadata, containing:*

*Version,*

*persistent ID,*

*ID for attribution,*

*anything referenced for the work, and*

*checksum (fixity) for establishing integrity of values.*

*The object may be data primarily, but should have attribution and provenance information. Provenance may include code/scripts used to derive or pre-process raw data in cases where code is relatively tightly bound to the data. If code is more generic/reusable, it may be better practice to cite code as a separate intellectual object (e.g., see Journal of Statistical Software).*

■ *We need ACM style guidelines for appropriate data citations: set of items of metadata*

*a.* *How to deal with old data? Never delete/replace anything, since data can be cited by others, and deleting/replacing can lead to inconsistencies. Need a versioning mechanism. Furthermore, citations should contain information to specify and verify the version used. See Principle 7 in the Joint Data Citation Principles.[[51]](#footnote-50)*

*b.* *Need support for retraction and updates of metadata*

■ *Develop guidelines for submission of data and use: develop workflow for integrating data.*

EPSRC “Expectation V

Research organisations will ensure that appropriately structured metadata describing the research data they hold is published (normally within 12 months of the data being generated) and made freely accessible on the internet; in each case the metadata must be sufficient to allow others to understand what research data exists, why, when and how it was generated, and how to access it. Where the research data referred to in the metadata is a digital object it is expected that the metadata will include use of a robust digital object identifier (For example as available through the DataCite organisation ‐ http://datacite.org).”

### Current practice

### Possible existing standards to adopt

The Cornell Research Data Management Group (RDMS) recommends the following components of a data citation:

creator(s) or contributor(s)

date of publication

title of dataset

publisher

identifier (e.g. Handle, ARK, Digital Object Identifier) or URL of source

version, when appropriate

date accessed, when appropriate[[52]](#footnote-51)

In addition, the site lists specific metadata standards that may be appropriate, though most are field-specific:

* + - * [Dublin Core](http://dublincore.org/) - domain agnostic, basic and widely used metadata standard
      * [DDI](http://www.ddialliance.org/) (Data Documentation Initiative) - common standard for social, behavioral and economic sciences, including survey data
      * [EML](http://knb.ecoinformatics.org/software/eml/) (Ecological Metadata Language) - specific for ecology disciplines
      * [ISO 19115](http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=53798) and [FGDC-CSDGM](http://www.fgdc.gov/metadata) (Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata) - for describing geospatial information
      * [MINSEQE](http://fged.org/projects/minseqe/) (MINimal information about high throughput SEQeuencing Experiments) - Genomics standard
      * [FITS](http://www.digitalpreservation.gov/formats/fdd/fdd000317.shtml) (Flexible Image Transport System) - Astronomy digital file standard that includes structured, embedded metadata

PANGAEA metadata includes

* Persistent identifiers (PIDs)
  + Digital Object Identifiers (DOIs),
  + accession numbers,
  + URLs.
* Relationships (i.e. DataCite schema)
* Provenance
* Author[[53]](#footnote-52)

### Micro data sets

Nature research journals have a policy of providing small data sets to readers accompany figures in the articles.

### Macro data sets

Larger data sets can be housed in recommended repositories with a minimum set of metadata

## Formats

What are recommended formats for submitted data?

The Cornell RDMSG recommends following the guidelines of the appropriate repository for choosing a data format, but offers three general principles:

* + 1. Select open, non-proprietary formats
    2. Select "lossless" formats
    3. Select unencrypted and uncompiled formats[[54]](#footnote-53)

## Review processes in place

Data quality review includes four areas, according to Peer et al:

1. “Review Files
   1. assign persistent identifiers
   2. create a citation to the study and a study-level metadata record
   3. record file details (size, format, checksums)
   4. Check that all files are present
   5. Verify that content of files matches expected format
   6. Create non-proprietary versions of the files
   7. Implement migration strategy for file formats
   8. Monitor bits
2. Review Data
   1. Check for undocumented variable and value information or out of range codes
   2. Review data for confidentially issues
3. Review Documentation
   1. Confirm comprehensive descriptive information for informed reuse including methodology and sampling information
   2. Link to other research products
4. Review Code

Check and verify code for data analysis and replication”[[55]](#footnote-54)

Elsevier data journals use a template and the following criteria for data review:

* Do the description and data make sense?
* Do the authors adequately explain the data’s utility to the community?
* Are the protocol/references for generating data adequate?
* Data format: is it standard for the field? Potentially re-usable?
* Does the article follow the required data article template?
* Is the data well documented?”[[56]](#footnote-55)

### Should review processes be the same or different for data and software?

### Branding Levels

#### Unreviewed, Verified, Validated, Checked, Certified, Replicable, Reproducible, Repeatable…

#### Available/Not available for distribution

Also see the [motivation](#h.yhvt4mgd1znl) section for best practices on encouraging authors and reviewers to participate.

The Research Data Alliance Sixth Plenary held a workshop on publishing data workflows.[[57]](#footnote-56),[[58]](#footnote-57)

## Storage recommendations/potential repositories

What are the storage recommendations?

■  *'Small' data sets could be stored in the ACM DL.*

What are the potential repositories to be considered?

■ *For bigger data sets, different communities have established their own repositories, e.g., Political Sciences use DataVerse, climate scientists use NOAA. For communities that do not maintain a repository, perhaps ACM could strike a deal with a big cloud company for storage, e.g., Google/Amazon/Microsoft. These companies already support free storage for academic/open data.*

■ *Many discipline-based repositories exist already: ICPR, data conservancy repositories, IEEE, etc. What about backup plans? What if we perceive one of these go away? Maybe have agreements in place for usage and recovery – get first right to harvest their data and reassign rights to ACM. ACM should draw up agreements with legal recourse.*

■ *ACM could be a repository of last resort for ACM publications, where no other disciplinary or institutional repository is suitable. The ACM Pubs Board could ratify criteria for external repositories -- looking at PLOS, etc. as a model. Suggest that minimal criteria include: independence of institution from single/small group of authors; commitment to persistence; creation of resolvable persistent id. Practically all credible archives / institutional repositories do this (including Dataverse, Figshare).*

■ *Digital preservation: think about risks. Should ACM create copies? Even if code lives on github, it is important to maintain the snapshot used to derive the results in the publications stored somewhere else. One option for this is zenodo -- see:*[*https://guides.github.com/activities/citable-code/*](https://guides.github.com/activities/citable-code/)

■ *Depending on the scale: keep provenance in the metadata. As long as large data sets are in “reasonably” long-lived institutions, there should be agreements to access the data.*

■ *Perhaps create an ACM “Data representation repository” that elevates it to a first class “archival” object.*

Potential repositories might include the following. A repository should demonstrate long term financial sustainability as well as the ability to integrate into ACM’s Digital Library.

* 3TU.Datacentrum
* ACM Digital Library
* Amazon
* Data Conservancy
* Dataverse
* Dryad
* Figshare

The American Chemical Society signed an agreement with Figshare in late 2015 where “Open access to supporting data [hosted at figshare] within any ACS journal will be freely available to all users via the Society’s publications website “[[59]](#footnote-58)

* Google
* Mendeley Data
* Microsoft
* PANGAEA
* Wikidata
* Institutional repositories
* Regional Repositories (including DataCite sites)
* Subject repositories

Nature’s *Scientific Data* journal maintains a list of recommended repositories in life sciences.[[60]](#footnote-59) There is also a directory of data repositories at re3data.org, the Registry of Research Data Repositories.[[61]](#footnote-60)

EPSRC “Expectation VII

Research organisations will ensure that EPSRC‐funded research data is securely preserved for a minimum of 10‐years from the date that any researcher ‘privileged access’ period expires or, if others have accessed the data, from last date on which access to the data was requested by a third party; all reasonable steps will be taken to ensure that publicly‐funded data is not held in any jurisdiction where the available legal safeguards provide lower levels of protection than are available in the UK.”

Also see EPSRC Expectation VIII and IX on data curation, noted under [Institutions](#h.5q10jr552t73) above.

A collaborative project of literature/data citation interlinking in beta by Data Literature Interlinking[[62]](#footnote-61) attempts to reduce the number of bilateral relationships between publishers, data centers, and bibliometric services.[[63]](#footnote-62)

## Citation

Should data be treated as a first-class object with its own identifier? Or as a supplemental file? If a supplement, should it have its own independent DOI and be independently citable?

Force11 has issued a set of data citation principles[[64]](#footnote-63) as has the Research Data Alliance.[[65]](#footnote-64)

The Sloan Foundation is funding a project called RMap which demonstrates the relationship between different objects in the research process, including articles, contributors, code, data, and text elements (like keywords) to name a few.[[66]](#footnote-65)

## Problems encountered

● What are the major problems encountered in submitting data and in using data?

■ *How do we deal with: Proprietary data? Old data formats/sets? Can these be updated/upgraded so that they becomes archival?*

■ *Develop guidelines for submission of data and use: develop workflow for integrating data.*

■ *Papers should contain a “data availability” statement, with a set of typical phrasings. (“The measurement data is available at …”, “The authors have declined to make the data available.”, “The data will be available on July 1, 2020 at …”)*

*a.* *To encourage authors to make data available, publications may allow deferred data availability. (This may address concerns about proprietary data as well as author concerns about being “scooped” for future work on the data set.)*

■ *Perhaps keep a working set of (domain/SIG-specific) data formats that are acceptable, so that authors can use these “standards” and reviewers don't have to deal with really old or one-off data sets. The list of domain-specific formats that can be revised occasionally by the SIG or journal (e.g., “input format for simulator X as specified in document Y”). This will allow datasets to be ingested in an easy manner.*

■ *Develop policies for risk management of data. How to deal with malicious data – where does ACM stand? What happens to the paper in case malicious data is discovered (e.g., human subject data obtained without IRB approval, copyright violation, plagiarism, pornography)? When do we require author consent? Should this be done by the DL, the journal’s Editor-in-Chief, the SIG leadership? Is there an appeals process for the aggrieved party?*

### Stability of Software used to access data

curate and preserve operating systems to support access to assets that depend on them

software applications

fonts, scripts, plug-ins and other dependencies

whole desktop environments to support access to the experience of interacting with it

preconfigured disk images with software already installed on them for running on emulated hardware

We need documentation, meaning persistent identifiers and catalogs for software, disk images and hardware for all this to work. The internet archive, NIST and Pronoun are working on these issues for software.

Emulation can be difficult to provide. One solution is Emulation as a Service (EAAS) provided to users remotely via a standard web browser.[[67]](#footnote-66),[[68]](#footnote-67)

Another set of documentation requirements:

* Dataset Description - <http://www.w3.org/2001/sw/hcls/notes/hcls-dataset/>
* Research Object - <http://www.researchobject.org/>
* Data Catalog - <http://www.w3.org/TR/vocab-dcat/>

### Proprietary Data

Copyright of collection and arrangement of data sets.[[69]](#footnote-68)

### Data Privacy

● Should a methodology section be mandated in any experimental paper?

■ *Methodology section should be mandated in experimental (including simulation) papers. But how should it be managed? Require a "Data and Methods” section. Methodology and data description can be included as an appendix but may not count towards page limits. Make it easy for authors. Organizers of domain-specific publications need to create policy and guidelines for data and methods. Charge SIGs to come up with specific policies, possibly with templates and examples provided by ACM or pioneer SIGs.*

■ *Joint declaration of data principles – with other professional societies, publishers…*

■ *Manage Data as structured objects – raw data, videos, etc.*

**Summary of Recommendations**

* Data should be a first class object, identified
* Data should be cited (in references) when used as evidence
* Citation should minimally include persistent ID, contributors, version
* Metadata should also minimally include fixity and references
* SIGs should develop domain-specific recommendations for data formats
* Data citation should resolve to an institution with a commitment to persistence
* Data citations should be verified before publication
* Papers should include a data availability statement
* Publications board should provide guidance on:
  + waivers for when data is not required/feasible
  + guidance for identifying institutions with commitment to persistence (and conversely, when copies must be replicated/deposited in ACM)
  + language/taxonomy for data availability statements (including statement of waiver)
  + acceptable restrictions on access
  + methods for corrections / retraction of data, including processes (who, why, appeal)

# Software

The software group looked at subject areas (by listing the Special Interest Groups) to determine:

* Requirements that are common for all SIGs or groups of SIGs
* Requirements that are unique

Our picture clearly showed clusters, but two common requirements did emerge: carefully documenting the experimental setup and the workflow of the experiment.

## Experimental Setup Template

The group felt that it having a standard template the CS research to document the experimental setup would be very useful to the CS research community. The group saw several benefits to using a standard template:

* While papers often include an experimental setup section, the section is terse and often leaves out many important details due to space constraints. A separate document (or supplement) would not be subject to space constraints and so it could be more complete and contain the necessary details. This document could be referenced from the paper thereby saving article space.
* Obviously, the use of the template would facilitate replicating and reproducing experiments.
* The group believed that a template could be developed that could be used by most research communities (of course, some sections might not apply and would be omitted). Using a standard template helps ensures important and necessary information is captured. It also makes it easier for reviewers to check, allows readers to find information quickly, and it allows necessary metadata to be automatically extracted for searching and indexing.

**Action Item:** Develop a prototype template and issue an RFC to the SIG communities.

## Workflow Template

In addition to the experimental setup, the workflow of the experiment must be recorded. Again, the group believed that a common template could be used.

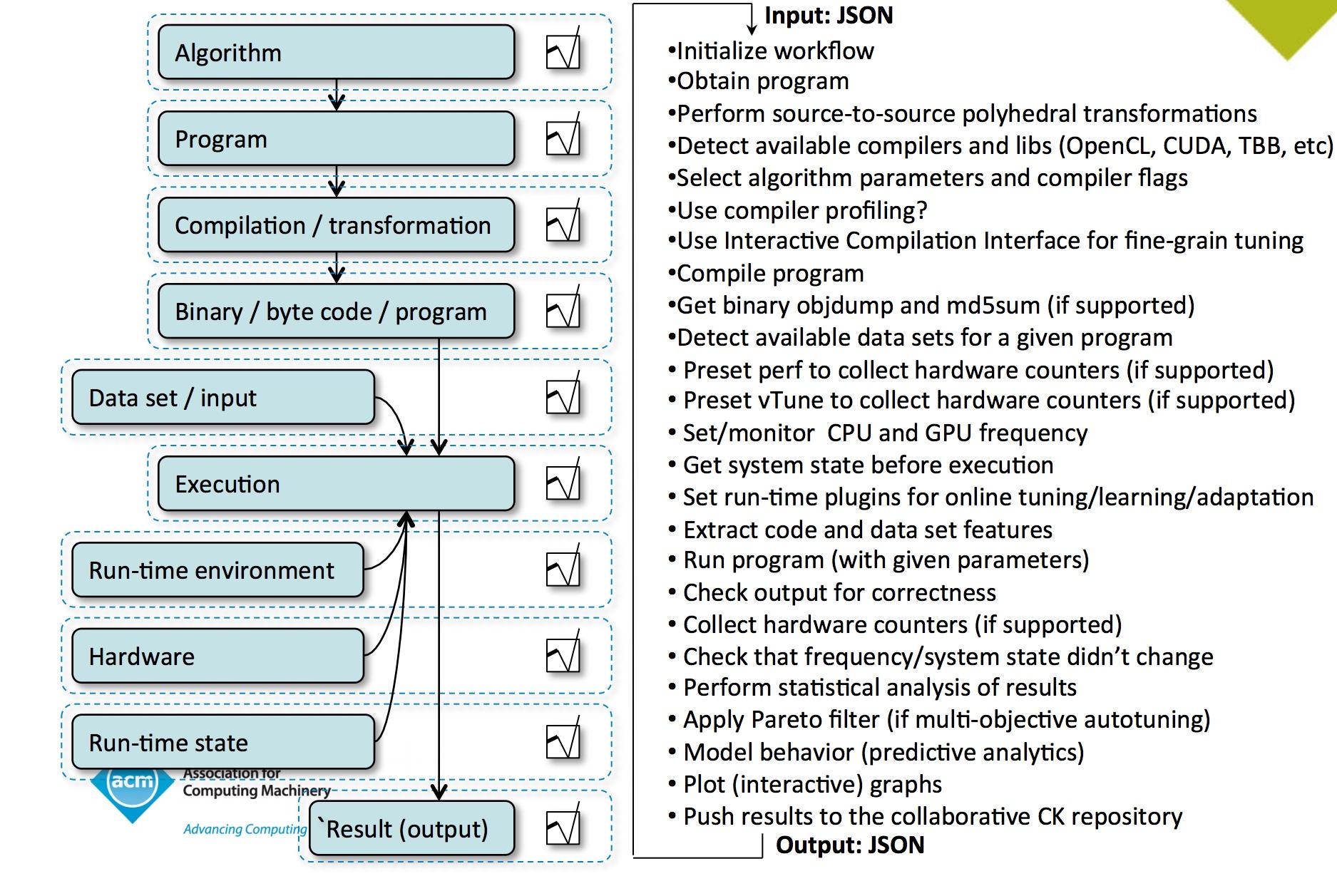
The benefits of the use of standard template are much the same as the use of a standard template for the experimental setup:

* Moves detail to a separate document saving article space and allowing automatic capture of information to facilitate indexing and searching
* Facilitates replicating and reproducing experiments

**Action Item:** Develop a prototype template and issue a request for comments (RFC) to the SIG communities.

### Workflow Template (Preliminary) Program software_workflow_template.jpg

Get real examples from different SIGS: for example, workflow in Collective Knowledge (CK) for software and hardware optimization and co-design <http://cknowledge.org/repo>



## Metadata

What Metadata description standards do we need to adopt or invent for software?

* Algorithms
* Software
* Benchmarks
* Environmental variables

JISC Software preservation ontology[[70]](#footnote-69)

Software Ontology (SWO)[[71]](#footnote-70)

Trustworthy Online Technical Environment Metadata (TOTEM) Registry[[72]](#footnote-71)

NIST National Software Reference Library[[73]](#footnote-72)

Preserving Virtual Worlds (FRBR-based ontology, serialized as METS or RDF-XML);

Challenges of determining what to include[[74]](#footnote-73)

Preserving .exe report, out of conference at Library of Congress

## **Identifiers for Software**

DataCite DOIs[[75]](#footnote-74)

[Based on git hooks](https://git-scm.com/book/en/v2/Customizing-Git-Git-Hooks)

Mozilla Science Lab -- Code as a Research Object

[Zenodo (CERN) DOIs for code in GitHub](http://www.software.ac.uk/blog/2015-07-28-making-code-citable-zenodo-and-github)

Should a methodology section be mandated in any experimental paper?

## c. Required parameters for short-term/long term reproducibility

Most of this section will be about being able to reproduce the software environment used in obtaining the original results. But another approach is creating emulators, as briefly mentioned in the Data section above. David Rosenthal of LOCKSS has surveyed the landscape of virtualization and emulators. He notes that two major hurdles exist; the inadequacy of existing tools and the legal status of emulating content.[[76]](#footnote-75) Nicole Contaxis discusses similar issues in her poster on decided how to preserve various pieces of software developed at the U.S. National Library of Medicine (NLM).[[77]](#footnote-76)

### Environment

Authors should explicitly specify the OS and tools that should be installed as the environment. Such specification should include dependencies with specific hardware features … or dependencies within the environment ...

### System

System setup is one of the most challenging aspects when repeating experiments. System setup will be easier to conduct if it is automatic rather than manual. Authors should test that the system they distribute can actually be installed in a new environment. The documentation should detail every step in system setup:

* How to obtain the system
* How to configure the environment if need be (e.g., environment variables, paths)
* How to compile the system (existing compilation options should be mentioned)
* How to use the system (What are the configuration options and parameters to the system?)
* How to make sure that the system is installed correctly

Ideally, the above questions should be easily answered by executing a set of scripts that will download needed components (systems, libraries), initialize the environment, check that software and hardware is compatible, and deploy the system.

SIGMOD Reproducibility committee recommends using the tool ReproZip.

### Experiments

“Given a system, the authors should provide the complete set of experiments to reproduce the paper's results. Typically, each experiment will consist of the following parts.

* A setup phase where parameters are configured and data is loaded.
* A running phase where a workload is applied and measurements are taken.
* A clean-up phase where the system is prepared to avoid interference with the next round of experiments.
* The authors should document (i) how to perform the setup, running and clean-up phases, and (ii) how to check that these phases complete as they should. The authors should document the expected effect of the setup phase (e.g., a cold file cache is enforced) and the different steps of the running phase, e.g., by documenting the combination of command line options used to run a given experiment script.

Ideally, experiments will be easier to reproduce if they are automatic, e.g., via a script that takes a range of values for each experiment parameter as arguments, rather than manual, e.g., via a script that must be edited so that a constant takes the value of a given experiment parameter.”[[78]](#footnote-77)

Manolescu and Manegold describe how to make experiments repeatable:

1. Making experiments portable
   1. not-so-exotic hardware
   2. free or commonly available tools
   3. scientific needs go first
2. Making experiments parameterizable
   1. credentials (OS, database, other)
   2. values of important environment variables (usually one or two)
   3. various paths and directories (see: environment variables)
   4. where the input comes from
   5. switches (pre-process, optimize, prune, materialize, plot . . .)
   6. where the output goes
   7. tricks
3. Building a test suite and scripts
4. Writing instructions
   1. What the installation requires; how to install
   2. For each experiment
      1. Extra installation if any
      2. Script to run
      3. Where to look for the graph
      4. How long it takes[[79]](#footnote-78)

### “Graphs and Plots

For each graph in the paper, the authors should describe how the graph is obtained from the experimental measurements. The submission should contain the scripts (or spreadsheets) that are used to generate the graphs. We strongly encourage authors to provide scripts for all their graphs using a tool such as Gnuplot or Matplotlib. Here are two useful tutorials for Gnuplot: a brief manual and tutorial, and a tutorial with details about creating eps figures and embed them using LaTeX and another two for Matplotlib: examples from SciPy, and a step-by-step tutorial discussing many features.

### Ideal Reproducibility Submission

The ideal reproducibility submission consists of a master script that:

* installs all systems needed,
* generates or fetches all needed input data,
* reruns all experiments and generates all results,
* generates all graphs and plots, and finally,
* recompiles the sources of the paper
* ... to produce a new PDF for the paper that contains the new graphs. It is possible!”[[80]](#footnote-79)

## d. Formats

What are recommended formats for submitted software?

## e. Review processes in place

“Check and verify code for data analysis and replication.” This means confirming the code executes and that it produces reported results.[[81]](#footnote-80)

Limor Peer presented the following guidelines for code review.

1. Before running the code

1. Do you have the data files the code calls for?

2. Set up working directories

3. Do you have the correct program and version?

4. What types of code are included?

Variable transformation and data cleaning

Statistical analysis

Presentation: tables, graphs, etc.

Saving, logging

5. Which commands do you have to run in order to replicate the analysis?

6. Is the code well-annotated?

B. Running the code

1. Open the published paper and identify results in the paper

Tables,

graphs,

in-text references

2. Run the code – does it fully execute?

3. Run the code – check against the published paper

C. Handling errors and discrepancies

Rule of thumb: no changes to PI’s code

1. Compile a list of values that do not replicate

2. Try and establish the reason for the error

3. Retrace your steps

4. If problem persists, follow repository policy.

### Should review processes be the same or different for data and software?

### Branding Levels

#### Unreviewed, Verified, Validated, Checked, Certified, Replicable, Reproducible, Repeatable…

#### Available/Not available for distribution

Also see the [motivation](#h.yhvt4mgd1znl) section for best practices on encouraging authors and reviewers to participate.

## f. Storage recommendations/potential repositories

What are the Storage recommendations? What are the potential repositories to be considered? The DL? Others? One? Many?

## g. Citation

Should software be treated as a first-class object with its own identifier? Or as a supplemental file(s)? If a supplement, should it have its own independent DOI and be independently citable?

## h. Problems

What are the major problems encountered in submitting software and in using software?

1. Rapid technological change
2. No archive of artifacts
3. Many implicit dependencies
4. Lack of backwards compatibility
5. Lack of social incentives**[[82]](#footnote-81)**

### Recreating the hardware and/or software environment

Early experience by several computing communities to introduce reproducibility reviews have uncovered a problem that the reviewer who is attempting to reproduce the results may not have access to proprietary, commercial, or specialized software and/or hardware used by the authors.

#### Solution:

Specify Hardware:

* CPU: Vendor, model, generation, clock speed, cache size(s)
* Main memory
* Disk (system): size & speed Software and Hardware Obsolescence

Specify Software:

* Product names, exact version numbers, and/or sources where obtained[[83]](#footnote-82)

“...the major problem … was a lack of performance portability in ever changing software and hardware, i.e. we could not get similar measurements even on very similar computers. So, the reproducibility issue in my domain [compiler and architecture design and optimization] is much deeper than just providing rules for packing artifacts, sharing them (as VM) and replicating outputs. We need to develop a strong evaluation methodology similar to physics to perform statistical analysis of all results (at least in our domain when dealing with stochastic behavior of computer systems) as well as to find missing “features” in experimental setups when results are not the same – it can be due to error, different system state (several other programs run at the same time, cache contentions, state of the network, etc).”

“Several authors ran their experiments on supercomputers with thousands of nodes or different GPUs and accelerator, or on hardware which is not yet in production, or using expensive proprietary software which reviewers may simply not have an access to.”

Two approaches are preserving the environment via virtual machines or Parrot or by cleaning up the environment using Umbrella or PRUNE.**[[84]](#footnote-83)**

### Relation of software artifacts to studies done with them in other sciences

### Lack of representative benchmarks and data sets

#### Solution:

“A public repository [http://cTuning.org](http://ctuning.org/) to let researchers share their benchmarks, data sets, and research tools along with publications in a unified format. At the same time, any user could spare computational resources to download shared experimental setups, perform some tuning and share back information about optimizations vs characteristics (crowd-tuning similar to SETI).”

### Large artifact sizes

#### Solutions

* Store in ACM Digital Library
* Store metadata in ACM DL and link to artifacts stored elsewhere

### Commercial interests of researchers to patent algorithms or software.[[85]](#footnote-84)

Also See [Competitive Concerns](#h.5lcv5iok7gkn) below.

## Review Processes

Publishers can promote the integrity of the research ecosystem by developing review processes that increase the likelihood that published experimental results can be independently replicated. An extreme approach would be to require independent replication as part of the refereeing process. An intermediate approach would be to require that artifacts associated with the work (e.g., data and/or software) undergo a formal audit.

In this report we outline terminology and minimal standards for review processes of these two types. Of course, there remain many circumstances in which such enhanced reviews will be either infeasible or not possible. As a result, we recommend that such review processes be encouraged, but remain optional, for journals and conferences, and that when they are made available, participation by authors also is optional. Authors who do agree to such additional review, and whose work meets minimal standards, should be rewarded with appropriate labeling both in the text of the article and in the metadata displayed in the Digital Library.

### Branding

We recommend that four separate reproducibility related brands be associated with research articles. Note that the three fundamental brands, *Artifacts Evaluated*, *Artifacts Available* and *Results Replicated*, are considered independent and any one, two or all three can be applied to any given paper depending on review procedures developed by the journal or conference.

#### *Artifacts Evaluated – Functional*

The artifacts associated with the research are found to be documented, complete, exercisable, and include appropriate evidence of validation.

##### Notes

*Documented:* An inventory of artifacts should be included, and sufficient description provided to enable the artifacts to be exercised.

*Complete:* To the extent possible, all components relevant to the article in question are included. (Proprietary artifacts need not be included. If they are required to exercise the package then this should be documented, along with instructions on how to obtain them. Proxies for proprietary data should be included so as to demonstrate the analysis.)

*Exercisable:* Included scripts and/or software can be successfully executed, and included data can be accessed and appropriately manipulated.

Artifacts need not be made publicly available to be considered for this brand. However, they do need to be made available to reviewers.

#### *Artifacts Evaluated – Reusable*

The artifacts associated with the article are of a quality that significantly exceeds minimal functionality. They are very carefully documented and well-structured to the extent that reuse and repurposing is facilitated. In particular, norms and standards of the research community for artifacts of this type are strictly adhered to.

##### Notes

Only one of the two *Artifacts Evaluated* brands should be applied.

Artifacts need not be made publicly available to be considered for this brand. However, they do need to be made available to reviewers.

#### *Artifacts Available*

Author-created artifacts relevant to this paper have been placed on a publically accessible archival repository. A DOI or link to this repository along with a unique identifier for the object is provided.

##### Notes

We do not mandate the use of specific repositories. Publisher repositories, institutional repositories, or open commercial repositories (e.g., figshare or Dryad) are acceptable. In all cases, repositories used to archive data should have a declared plan to enable permanent accessibility. Personal web pages are not acceptable for this purpose.

Artifacts do not need to have been formally evaluated in order for an article to receive this brand. In addition, they need not be complete in the sense described above. They simply need to be relevant to the study and add value beyond the text in the article. Such artifacts could be something as simple as the data from which the figures are drawn, or as complex as a complete software system under study.

#### *Results Replicated*

The main conclusions of the paper have been independently verified by a person or team other than the authors.

It is easy to see how research articles that develop algorithms or software systems could be branded as described above. Here, the artifacts could be implementations of algorithms or complete software systems, and replication would involve exercise of software, typically software provided by the author. However, we intend these brands to be applicable to other types of research as well. For example, artifacts associated with human-subject studies of novel human-computer interface modalities might be the collected data, as well as the scripts developed to analyze the data. “Replication” might focus on a careful inspection of the experimental protocol along with independent analysis of the collected data.

### Review Procedures

The descriptions of brands provided above do not specify the details of the review process itself. For example: Should reviews occur before or after acceptance of a paper? How many reviewers should there be? Should the reviewers be anonymous, or should they be allowed to interact openly with the authors? How should artifacts be packaged for review? What specific metrics should be used to assess quality? Current grassroots efforts to evaluate artifacts and formally test replicability have answered these questions in different ways. We believe that it is too early to establish more specific guidelines for artifact and replicability review. Indeed, there is sufficient diversity among the various communities in the computing field that this may not be desirable at all. We do believe that the broad definitions provided above provide a framework that will allow brands to have general comparability among communities.

We acknowledge that the brands described above make sense even if they result from an evaluation that occurs after publication. Editors-in-Chief and Program Committee Chairs should have the authority to award any of the brands listed above post-publication if warranted.

### Caveat Regarding Distribution of Binaries

Professional societies, and publishers in general, need to be wary of distributing binaries provided by authors. Even when authors are making good efforts to keep their software clean, binaries can include malware or proprietary components. While publishers can protect themselves to an extent through the use of release forms, the reputation of an organization can be seriously harmed if binaries they distribute cause harm to systems of referees or digital library users.

# Publishing Policies

The American Astronomical Society has a policy explicitly promoting the citation of software associated with published research. The policy also encourages the publication of papers describing research software[[86]](#footnote-85).

Emilie Marcus of Cell Press describes a number of actions publishers are taking to increase reproducibility:

1. Reporting
   1. Improving methods, setting standards, including expanding methods section, tagging methods so they are discoverable, embrace agreed-upon guidelines in fields where they work, and creating and supporting methods-only journals to allow the space and attention to documenting procedures.
   2. Access to reagents and research materials
   3. Protocol methods/repositories
   4. Improving reporting of statistical information
   5. Open access to analytical methods, code, software
   6. Registered reports or preregistered experimental methods and analysis.
2. Sharing (data sharing)

Marcus notes, “Just sticking data up somewhere isn’t useful to anyone.” Increased vigilance in screening, setting peer review standards and searching for reviewers, and post-publication corrections and evaluation are all important.

Publishers can promote accountability by publishing replication studies and supporting contributorship identification such as the CRediT badges.[[87]](#footnote-86)

Andrew Hufton from the journal *Scientific Data* at Nature Publishing Group has described a number of steps the publisher has taken to “reduce irreproducibility.”

1. Reporting Checklist

Nature has created a reporting checklist[[88]](#footnote-87) for key methodology issues as a tool for the editors. The checklist is not published. It has also been adopted by BioMed Central (BMC) journals, and has been endorsed by the US National Library of Medicine (NLM).

2. Data deposition and sharing.

Though Nature has had a longstanding “share upon request policy,” it can be limiting. Publishers need to remove barriers to sharing. Nature titles explicitly allow pre-publication sharing and article preprints; such pre-publication sharing will not compromise publication of articles. Publication of data articles without results or analysis will also not compromise publishing decisions. Similar policies have been adopted at the BMC journals.

For small data sets, the journals have ways to provide the data behind plots and graphics to readers.

The journals are enforcing data deposition rules. *Scientific Data* maintains a list of recommended public data repositories for the life sciences.

In order to ensure scientists get credit for sharing their data, *Scientific Data* publishes articles focused on data reuse. They are published with a Creative Commons Attribution license (cc-by). Data quality is as important as the reliability of results. They attempt to eliminate walled community data gardens. The articles link to a data store in a community repository.[[89]](#footnote-88)

## Reproducibility “Label” or Badge

## Ethics & Plagiarism

## Privacy Concerns/IRB issues (see [data](#h.hbgu63gkizgc)?)

## Competitive Concerns

EPSRC “Expectation VI

Where access to the data is restricted the published metadata should also give the reason and summarise the conditions which must be satisfied for access to be granted. For example ‘commercially confidential’ data, in which a business organisation has a legitimate interest, might be made available to others subject to a suitable legally enforceable non‐disclosure agreement.”

Disclosures (pre- and post-acceptance)

## Works-in-progress versus archival publications

## Genre distinctions

## Create a Separate Dataset Track

## Publicize evolving list of problems with solutions[[90]](#footnote-89)

## Required vs. Optional

## Actions if the reproducibility review fails

This blog post by a political science, may help further clarify the issues when a reproducibility review fails:

"...when a specific input fails to produce the originally described output, actions should be taken. But the type of action taken depends on what steps of the scientific process (i.e., what inputs) are being recreated... There is no single reasonable response to a "replication failure" and the implications of replicability and reproducibility are quite distinct. We need to be clear - regardless of what terminology we use - about what output is being recreated (or failing to be recreated) from what input when we consider why reproducibility or replicability are important features of science."[[91]](#footnote-90)

## Preservation/Long term Storage

NISO recommends the format should be supported by open source software.

# Implementation Issues

## Overview

From the publishing perspective, our main challenge is to integrate the outcomes of this Task Force and its sub-groups into the publications process and into the final Digital Library presentation of artifacts as stand-alone primary objects and in association with research articles and experiments.

To date, the experiments run by many of the ACM conferences and a few of the ACM journals are divorced from the existing workflows for manuscript submission and review that exist today.

In order to scale these operations, we need

* To implement a systematic way to handle deposit and review of artifacts
* To integrate data and software repository systems with existing manuscript submission workflows, while allowing for anonymity and security, and to integrate the ACM Digital Library with the chosen external repositories
* To create new or enhance our existing bibliographic XML DTDs to capture standardized metadata for artifacts and figure out the proper way to identify and display them in the Digital Library with their associated level of branded review
* Standard citation and linking mechanisms for artifacts
* Standard reference citation formats for all artifact types

## Requirements for Implementation

Our group identified what we need to get from the other groups to implement a publishing program that enables the replication of research results in the computing field and supports review and publication of artifacts.

1. **Metadata** definition standards (at multiple levels)

a. What are the types of artifacts and what are their metadata schemas/vocabularies

b. What standards exist that can be adopted

c. How to model a collection of artifacts (e.g. via OAI-ORE)

d. Names and descriptions of review types

e. Licensing and rights (including compliance with all appropriate legal/privacy/ethics laws)

2. **Ethical Best Practices** (We expect part of the guidance here to be provided by the ACM Committee on Ethics and Plagiarism)

a. For authors regarding the collection of data, anonymizing data, code commenting, plagiarism, how to credit and cite software

b. Recommendations for licensing and access, re-use rights, redistribution

3. **Granularity** issues

a. Granularity of identifiers; when is an artifact identified with its own DOI; when is it a first-order independent object; when is it a sub-component of an article; or a collection of objects; how are the various relationships specified

b. Differentiation between containers (e.g. in the VM sense) and collections (in the more abstract sense, eg. ORE)

c. Multiple container definition approaches (ORE, BagIt, etc.)

d. What bits are citable?

**4.** **Versioning**

a. Should ACM support software and/or data set versioning (or rely on external specialized repositories for this)?

b. Should versioning be supported post-publication?

c. Are older versions maintained or treated as (annotated) replacement updates?

5. **Review** standards

Need to understand the various levels and definitions of review and their workflows (for submission tracking systems and for branding)

**6.** **Repositories**

a. *Assumption #1*

ACM will need to allow for integration with external repositories (rather than to assume that all artifacts will go into the ACM Digital Library or a new ACM managed repository). Advice will be needed on recommended repositories. The group identified the following criteria for external repositories:

i. Integration with manuscript submission

ii. Developer community robustness

iii. Longevity & egress SLAs

b. *Assumption #2* - ACM will also need the Digital Library to support artifacts or build their own default repository (for this and other use cases)

i. Long-term archiving/preservation and access

ii. Dark archiving use cases (privacy, confidentiality)

c. Criteria for selecting from existing repository software packages versus building

i. Developer community robustness

ii. Features (e.g. flexibility of features, ability)

iii. Security

iv. Versioning

## c. Submission Systems-Repository integration

Integration of manuscript submission systems in use with a data or software repository

## d. Implications for authoring templates

## e. Citation and Identification

Identification and linking of the articles and their related data and software

### DOI assignment –

### first class objects/components

## f. Branding

What gets branded, and how and where that brand is displayed?

• Reviewed/Unreviewed

• Distributed/Not distributed

## g. Promotion of reproducibility

### Authors

### Editors

### Publishers

## h. Document Structure

Research Objects[[92]](#footnote-91)

## i. Linking procedures

## j. Bit Rot

## k. Tools

* + - Docker[[93]](#footnote-92)
    - iPython Notebook[[94]](#footnote-93)
    - FairDom tools[[95]](#footnote-94)
    - myGrid tools[[96]](#footnote-95)

SEEK[[97]](#footnote-96)

# Appendix A. Background Experience

This appendix summarizes the experience of each of the participating communities with regard to reproducibility. These experiences were drawn on to formulate the best practices guidelines in the main body of the document.

## ACM Special Interest Group Governing Board (SGB)

The [ACM](http://www.acm.org/) [SGB](http://www.acm.org/sigs/sgb/) Replication Task Force was initiated at the request of the SGB Chair Patrick [Madden](http://dl.acm.org/author_page.cfm?id=81100092346) in September 2015. SIG Development Advisor [Simon Harper](http://dl.acm.org/author_page.cfm?id=81100139139) was appointed Taskforce Chair to take this forward as well as integrating it into the 'DL-Technology: Data, Software, & Reproducibility' project.

The charge of the task force is to develop proposals on how ACM can bring current replication and verification practices in line with the rest of the scientific community. Specifically, the following questions are of interest.

1. Across the SIGs, what mechanisms are in place to enable replication and verification efforts? Are they sufficient to catch errors or deliberate fraud (and have they in fact done so?). Can these processes be improved or encouraged in some way?
2. What are reasonable expectations from authors, with regards to enabling replication and verification? How can these expectations be made compatible with the review process in conferences, symposia, journals? Should there be different expectations during review, and after paper acceptance?
3. If published work fails to be replicated, appears to contain significant errors, or even appears to be fraudulent — what mechanisms should be in place to investigate the matter? Who would such concerns be communicated to? Who would perform investigation, and decide the matter? And if errors, deception, or fraud are confirmed — what should the ACM, as a publisher, do? Should there be corrections, annotations noting the errors, paper retraction?

The SIGs, with elected leaders representing the various research communities of computer science, are a logical group to engage in setting policy. The SIG leaders will be familiar with typical practices within their areas of expertise. With control of many of the major publishing venues, this group also has the ability to enact change.

We have an open and transparent document evolving at:

<https://github.com/sharpic/Replication-Taskforce>

## Special Interest Group on Management of Data (SIGMOD)

**Stratos Idreos**

The text below is taken from the SIGMOD web site for reproducibility for database research.

“**What is SIGMOD Reproducibility?**

SIGMOD Reproducibility has three goals:

* Highlight the impact of database research papers.
* Enable easy dissemination of research results.
* Enable easy sharing of code and experimentation set-ups.

In short, the goal is to assist in building a culture where sharing *results*, *code*, and *scripts* of database research is the norm rather than an exception. The challenge is to do this efficiently, which means building technical expertise on how to do better research via creating repeatable and sharable research. The [SIGMOD Reproducibility committee](http://db-reproducibility.seas.harvard.edu/#Committee) is here to help you with this.”

The web site contains materials about definitions, processes, tutorials, motivation, guidelines, etc. that are may be useful as an example . It is a voluntary post-acceptance process

## MultiMedia (MM)

We identify two aspects in the reproducibility of MultiMedia (MM)-oriented scientific results: dataset and code. On both aspects, the Multimedia community has some actions:

* Dataset: In general, the MM community uses "[representative datasets](http://www.ee.cityu.edu.hk/~lmpo/lenna/Lenna97.html)". Since 2011, the [dataset track at MMSys](http://www.mmsys.org/index.php/mmsys-datasets) is a way to select and highlight these datasets. Some of MMSys datasets have been well cited. The use of representative datasets is not directly linked to reproducibility but at least it helps scientists to compare their proposals. Here are some of the most prominent datasets:
  + [MediaEval](http://www.multimediaeval.org/about/): a framework for testing algorithms
  + [Qualinet](http://dbq.multimediatech.cz/): a focus on Quality of Experience (QoE)
  + [FlickR database](http://labs.yahoo.com/news/yfcc100m/): million of images for research
* Algorithm code: The [Open Source group](http://sigmm.org/Resources/software) at SIGMM has been active for years. In particular, this group organize the annual [open source competition](http://sigmm.org/Resources/software/ossc) at the flagship ACM MM conference. It promotes scientists who publicly release research-oriented code. The selected projects got a citation in the MM conference and appear as publications in ACM DL.
* Other actions:
  + Also associated with the ACM MM conference is the [Grand Challenge](http://www.acmmm.org/2015/call-for-contributions/multimedia-grand-challenges/) competition, which is initiated by companies and aims at making scientists focus on industry-oriented issues. The challenges usually include a dataset, and the all the contributions from scientists are expected to be reproducible.
  + A minor but significant action was to include the question "Does the paper provide public dataset?" in the review form of the ACM MM’15 conference.

The SIGMM and image processing researchers have a long experience of using the same image/video to test their algorithms. A funny consequence of it is that it enables community-sharing stories. See for instance:

<http://www.ee.cityu.edu.hk/~lmpo/lenna/Lenna97.html>

Since 2011, the multimedia system community (MMSys, which is a part of SIGMM) has offered a "dataset track" in the main MMSys conference.

<http://www.mmsys.org/index.php/mmsys-datasets>

The goal is not as ambitious as the SIGMOD reproducible label: "We encourage the use of common datasets for experimentation. This makes it easier to compare results from different research groups. The dataset track at the conference allows contributors to receive some credit and citations for the datasets that they have contributed."

We provide a repository. The papers that are associated with the dataset, appear in the MMSys proceedings as for any other paper. Some papers have been well cited, for example 97 citations for a 2012 paper:

[https://scholar.google.com/citations?view\_op=view\_citation&hl=en&us…](https://scholar.google.com/citations?view_op=view_citation&hl=en&user=WqVnh0IAAAAJ&citation_for_view=WqVnh0IAAAAJ:SP6oXDckpogC)

I can try to gather more details about this dataset track for our workshop (including the number of submitted papers and an accurate record of citations).

## ACM Transactions on Mathematical Software (TOMS)

### Software Exchange in the Mathematical Software Community

**Ron Boisvert**

Since its inception in 1975, the *ACM Transactions on Mathematical Software* *(TOMS)* has published software in association with a subset of its articles. This activity is rooted in the early practice of publishing “algorithms” within the pages of the *Communications of the ACM* (CACM). These contributions were numbered, with the series known as the *Collected Algorithms of the ACM (CALGO)*. This practice was discontinued in 1975, and formally transferred to *TOMS*. The first *TOMS* algorithm was numbered 493.[[98]](#footnote-97) Instead of printing the algorithms, the software was archived and made available via a distribution service (initially on tape, but later through the web). Today, machine-readable versions of all *TOMS* algorithms are available as supplementary material in the ACM Digital Library. The most recent algorithm, published in June 2015, is numbered 950.[[99]](#footnote-98) In addition, some of the original *TOMS* has been able to maintain a steady flow of algorithms (roughly 2-3 per issue) for 40 years.

## *IEEE/ACM Transactions on Networking* (TON)

**Henning Schulzrinne**

The *IEEE/ACM Trans Networking* allows authors to submit software and data as supplementary materials but does not require them to do so.[[100]](#footnote-99) The Steering Committee has found requiring authors to submit their data problematic for several reasons. First, much of the data associated with its published articles is considered proprietary (e.g., by Internet service providers or cloud service providers). Another issue is data privacy. Some articles reference data derived from experiments that require Institutional Review Board (IRB) approval, especially when there are human subjects. Authors are concerned that data could be de-anonymized; many find the risks to be greater than the reward.

## IEEE Signal Processing Society

In 2005 the IEEE Signal Processing Society, in the field of computational science, attempted to reproduce results from papers published in 2004 the *IEEE Transactions on Signal Processing.* The results were discouraging, and not much of the material was shared. Though IEEE is interested in promoting reproducibility, the experiment was not terribly successful, because it was voluntary with no infrastructure. IEEE Xplor allows authors to upload supplemental materials, but this feature is underutilized. There should be a standardized way to present data, and that should be promoted whenever possible. It is important to share code, data, and algorithms, but also information about the computing environment. One publisher or organization should not create its own proprietary approach. A public-private partnership with funding agencies and publishers might be helpful. The culture must be changed to increase sharing, but this is difficult to change. Funding agencies such as the National Institute of Health (NIH) and the National Science Foundation (NSF) are likely to begin requiring open data, just as they required open access, as indicated in a February 2015 National Academy of Sciences Workshop on reproducibility. Multiple parties have roles to play: funding agencies, journal editors-in-chief, scientific societies, libraries, publishers, and universities.[[101]](#footnote-100)

## Supercomputing (SC)

Reproducibility in supercomputing has long been an issue, as discussed by Bailey as early as 1991.[[102]](#footnote-101) In fact, there is an interesting Annual Workshop on Duplicating, Deconstructing, and Debunking, where the papers that were presented at the International Symposium of Computer Architecture (ISCA) the previous year are dissected. An example from this year examines the runtime scalability of a popular benchmark suite.[[103]](#footnote-102)

“Experimental reproducibility is a cornerstone of the scientific method. As computing has grown into a powerful tool for scientific inquiry, computational reproducibility has been one of the core assumptions underlying scientific computing. With "traditional" single-core CPUs, documenting a numerical result was relatively straightforward. However, hardware developments over the past several decades have made it almost impossible to ensure computational reproducibility or to even fully document a computation without incurring a severe loss of performance. This loss of reproducibility started with CPUs that used out-of-order execution to improve performance. It has accelerated with recent architectural trends towards platforms with increasingly large numbers of processing elements, namely multi core CPUs and compute accelerators (GPUs, Intel Xeon Phi, FPGAs).

Programmers targeting these platforms rely on tools and libraries to produce codes or execute them efficiently. As a result, codes can run efficiently, but have execution details that can be impossible to predict and are often very difficult to understand after execution. Furthermore, parallel implementations often result in code with varying execution orders between runs, leading to non-reproducible computations. The underlying reasons are that (1) the hardware and system software allocate parallel work in ways that are not always specifiable at compile time and (2) the execution often proceeds in an opportunistic manner with the execution order changing between runs.

As such, floating-point computations, which are not commutative and associative, can have different execution orders and execute on different processing elements between runs, leading to runs with varying results as a matter of fact. The predictability of systems is further complicated by two issues that are becoming more critical as systems grow in scale: (1) interconnect systems with latencies that are often outside the control of programmers and (2) reliability as the mean time between failure (MTBF) is now measured in hours on large systems. This situation seriously affects the ability to rely on scientific computations as a metrological substitute for experimentation!”[[104]](#footnote-103)

### SC16

**Michela Taufer**

The following is an update on activities from SC16 - The International Conference for High Performance Computing, Networking, Storage and Analysis ([http://sc16.supercomputing.org](http://sc16.supercomputing.org/)).

SC16 is the leading conference in High Performance Computing (HPC).

The Steering Committee of the conference is supporting an initiative that aims to study how to integrate aspects of research reproducibility, reusability, and sharing of codes and data within the technical program of the conference. An ad-hoc committee is working on strategies to develop an instruction manual that support “research reproducibility” where “reproducible research” and “reproducibility” refers to the ability to recreate computational and/or scientific results from the data and code used by the original researcher.

We borrowed the definition of “research reproducibility” from these two sources:

* ICERM Report: <https://icerm.brown.edu/tw12-5-rcem/icerm_report.pdf>
* Biostatistics Editorial: <http://biostatistics.oxfordjournals.org/content/10/3/405.full.pdf+html>

SC16 is using a part of its program called Student Cluster Competition (SCC) to create a prototype of an instruction manual for papers. The Student Cluster Competition (SCC) is designed to introduce the next generation of students to the high-performance computing community. Teams of undergraduate and/or high school students assemble a small cluster on the SC16 exhibit floor and race to complete a real-world workload across a series of applications and impress HPC industry judges.

SC16 is inviting authors of technical papers accepted at past SC conferences to submit their paper’s codes and data that can be transformed into applications for the SCC. Papers selected as SCC applications will be supplemented with an instruction manual that includes rigorous documentation of the experimental environment, the methodology for experimental replication, and the validated outcomes (results and/or performance). Eventually a version of the tuned instruction manual could be considered as a required companion of any paper submitted to the conference.

We want to share a set of slides that was presented at the Steering Committee meeting last week in which the ad-hoc committee reported on the initiatives of other communities as well as the work at SC16.

## Artifact Evaluation (Compilers/Software Engineering)

Most of the *TOMS* algorithms are implementations of new methods for solving generic mathematical problems that recur in applications, such as evaluating mathematical functions, solving linear systems of equations, and finding the minimum of a multivariate function. In these cases it is natural to package the software as a reusable component; hence the emphasis on portability in the evaluation. Software distributed by *TOMS* is typically packaged as a file bundle, such as a zip archive, containing the component of interest, along with test drivers, test data, documentation, etc.

A good deal of success has been achieved in keeping all of the *TOMS* algorithms back to Algorithm 493 usable today. Part of this has been the emphasis on using widely available computer languages and standard constructs. Nevertheless, some have aged better than others. In 2002 *TOMS* Algorithms Editor Tim Hopkins systematically evaluated the *TOMS* algorithms to date, making a variety of adjustments needed to make them usable.[[105]](#footnote-104)

### Encouraging Replicability of Results in ACM *TOMS*

**Ron Boisvert and Michael Heroux**

The *ACM Transactions on Mathematical Software (TOMS)* has recently initiated a formal process to assess replicability of results reported in its published articles. The TOMS RCR initiative introduces the weakest form of reproducibility: One person certifies that computational result are reproducible at one point in time. We do not attempt to formally archive artifacts for the future. Also, the review is optional at this time.

We have a specific reviewer whose sole responsibility is to perform the RCR review and describe the process used. This person's report is an official ACM TOMS publication, and additional artifacts can be added to the digital library providing detailed evidence from the review process.

Although the RCR process is very minimal, it has already had an upstream positive impact on practices for authors who wish to take part in the review. We have heard from them about improved management and versioning of software and data, and more care in generating results. To me, this is the best impact of the effort. It's not about catching bad results as much as signaling the importance of good results and seeing the improved upstream behavior. The process is described in detail in an editorial written by *TOMS* Michael Heroux[[106]](#footnote-105).

Some of the features of this initiative are described below.

* The Replicated Computational Results (RCR) process is an optional additional review process initiated when the Editor-in-Chief determines that the article has a reasonable chance to be published, e.g., after the first round of reviews.
* The Editor solicits a single additional reviewer to perform the RCR review. It is expected that the RCR reviewers will typically be senior graduate students or postdocs. The main job of this reviewer will be to determine whether the main results of the paper can be replicated. Ideally, this is done by some form of computational replication. The identity of the RCR reviewer is known to the authors, in order to expedite the resolution of problems discovered during the replication process. In some cases, for example, it may be necessary for the reviewer to obtain access to specialized computing facilities provided by the author to carry out the work.
* When the process is complete the RCR reviewer writes a sort (e.g., 4-5 page) report describing the replication study, any issues discovered along the way which would be a help to others contemplating similar work, as well as the overall conclusions. This paper is published along with the paper being reviewed in the same issue in which the paper appears. In this way, the reviewer receives tangible credit for undertaking the replication study.
* If the results of the paper are deemed replicable, a notation to this effect is placed with the paper itself, as well as on the article’s citation page in the ACM Digital Library.

The *TOMS* RCR process was piloted before the announcement, and the first paper receiving the RCR designation appeared in the same *TOMS* issue as the editorial above.[[107]](#footnote-106) [[108]](#footnote-107)

### Artifact Evaluation Process Bruce Childers (University of Pittsburgh), Grigori Fursin (cTuning Foundation)

Artifact Evaluation (<http://www.artifact-eval.org>, AE for short) originally grew out of concern for the insufficient attention on the artifacts that back experimental results in compiler/programming language and software engineering research. Other communities in computer systems face a similar situation.

AE affords an opportunity for researchers to submit for evaluation artifacts and experiments that accompany their papers. The goal is to reward authors who make an effort to create useful artifacts beyond the paper, and to construct quality experiments, which they release to the broader community. The process aims to improve accountability to enable repeatable and reproducible experiments and to leverage building directly upon other researchers’ artifacts and experiments.

AE rewards robust, well documented and validated artifacts and carrying out sound experiments to encourage accountable experimentation. While details of AE can be tailored, the general principles are similar for authors that decide to participate in AE.

1. First, authors prepare, package and deliver artifacts and associated experiments to an evaluation committee (EC), which may be separate from the program committee (PC).
2. Second, the EC examines the artifacts and experiments to determine if they meet community-determined standards of “validated”.
3. Third, authors are notified whether their artifact and experiments were validated.
4. An optional rebuttal step may also be used prior to making final validation decisions.
5. Finally, authors are rewarded for meeting the standard.

A few variants of the process are possible. In step 1 above, although it could be made compulsory, this is usually viewed as too aggressive, imposing too high of a barrier to entry (e.g., excluding papers using proprietary or commercial software, hardware, or data sets).

Steps 3 and 4 may be concurrent, post, or two-phase with paper review. With a concurrent process, AE can influence paper review decisions. However, a concurrent process can impose a large collective burden:

(a) because evaluators may have to review hundreds of artifacts; (b) authors of rejected papers may be discouraged.

In a post-acceptance process, only artifacts and experiments from accepted papers are evaluated: (a) reduces the collective effort; (b) separates paper evaluation from the artifact, ensuring the paper stands on its own, as an archival result; (c) raises issues when evaluators find problems with the artifacts or experiments for an accepted paper (e.g., should the paper be re-evaluated?).

In the 2-phase process, AE is carried out for a larger subset of papers, and the validated ones are accepted, thus creating a higher bar for acceptance. Another important choice is what is the standard to be “above threshold”? For example, should an experiment be entirely open and constructed from scratch to be validated? Typically, evaluation committees take some middle ground on evaluating an artifact and experiments, focusing on best faith effort.

In the last step, reward types also vary:

(a) simple recognition of a job well-done, through “reward seals” or a public list of validated artifacts;

(b) additional time to present the work (perhaps in a separate paper track;

(c) additional pages in proceedings;

(d) “best artifact” awards;

(e) financial rewards/awards.

#### Experience with AE at CGO’15 and PPoPP’15

As examples of the choices, we highlight our recent experience with the process. We co-organized AE for the IEEE/ACM Symposium on Code Generation and Optimization (CGO), 2015 and the ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP); the conferences jointly used AE as a post-acceptance, opt-in process with a rebuttal step. Authors could use any “reasonable” means to provide access to artifacts and experiments.

The EC (different from the PC) was composed of senior graduate students and postdocs to give them experience in AE, and also because they are most familiar with recent technologies embodied in the artifacts and experiments. Authors decided how much to provide, e.g., do they include source code, all data sets, benchmarks. Both conferences rewarded artifacts by listing those that were validated. For example, CGO and PPoPP 2015 distributed a seal for authors to add to their papers, presentations and/or websites.[[109]](#footnote-108) The seal also appeared in the ACM Digital Library. 

Note that since AE is new to compiler and architecture conferences, it can expose problems that have not been envisioned in other related and already established AE initiatives such as for machine learning conferences. Therefore, rather than enforcing specific rules for packing and replicating/reproducing experiments, we let authors decide what to include and how to package their artifacts together with a “Howto” documentation to explain building, using and running their artifacts to reproduce experiments.

In general, authors expressed a strong interest in participating in the process. The participation rate was approximately 40% between the two conferences. Through the experience, we learned a few lessons.

First, there can be confusion by what “validated” means by both authors and evaluators. Authors and evaluators were confused whether the experiment or the artifact was the object of the evaluation. This was a particular issue when authors assumed the focus was the experiment, and then the evaluator assumed it was the artifact.

Second, authors used many different mechanisms to package and provide access to their artifacts, including compressed tar balls, VMWare images, CDE, Docker and Virtual Box virtual machine. The latter was the most popular mechanism.

Third, there was some contention over access to proprietary, commercial, or specialized software and/or hardware needed to evaluate an experiment. For example, one paper relied on custom power management hardware; in this case, the authors provided raw data traces from their instrument. In another example, the authors used proprietary tools in their evaluation and they provided remote access to a machine that had the tools available, satisfying licensing concerns.

We found that the mechanism for AE packaging and delivery was every bit as critical as the process itself. Since we have to deal with ever changing hardware and software, even tiny changes in experimental setup can cause failure of Virtual Machine or Docker images. On the other hand, rebuilding software for a given hardware from scratch often fails due to missing dependencies. Furthermore, even if we manage to replicate and run an experiment, it is often impossible to reproduce performance or energy measurements due to highly stochastic behavior of computer systems and lack of run-time state. Therefore, we will need to develop a common methodology for statistical analysis of experimental results (for example, similar to physics or electronics) as well as mechanisms to describe how to rebuild/rerun experiments with all software, hardware and run-time state dependencies. These problems in our domain put extra burden on authors and evaluators. Indeed, evaluators in the CGO and PPoPP EC did more than one AE, like in a program committee, and the evaluators had to learn how to evaluate (unpackage, run, etc.) each artifact/experiment, which sometimes contained incomplete directions or missing files (by mistake or proprietary rules).

Finally, since our longer-term goal is to improve AE process and establish a common methodology via common discussions with the community, we also allowed communication between authors and reviewers during AE to solve obvious problems with installing and replaying experiments. However, this raised a privacy issue - although authors of the papers participating in AE are no longer anonymous (since the paper is already accepted), the reviewers must stay anonymous, which is particularly important in case of rejected artifacts, which may cause hard feelings. Therefore, we decided to serve as a proxy between authors and reviewers. This process turned out to be rather inconvenient and unscalable. We envision several possible solutions: either disallowing such communication, or using anonymous and closed discussion websites, or making AE process completely open allowing anyone participate in evaluation and discussions while letting AE committee make the final decision.[[110]](#footnote-109) [[111]](#footnote-110)

#### Experience with AE at CGO’16 and PPoPP’16

New guidelines were developed and used for the 2016 conference based on experience in 2015 and discussions among the ACM Digital Library reproducibility working group held in July 2015.

[http://cTuning.org/ae](http://ctuning.org/ae)

[http://cTuning.org/ae/submission\_extra.html](http://ctuning.org/ae/submission_extra.html)

* We provided a LaTeX template to describe submitted artifacts and experiment workflows with the informal meta information - we plan to use it to derive some common keywords for the ACM DL at least across computer systems' conferences.
* We added version to submission/reviewing templates to keep track of the methodology.
* We plan to let authors keep this AE template as Appendix in the final paper together with the template version - it should allow readers to trace back the reviewing procedure.

##### Results

* 23 artifacts were submitted in 2016 vs. 17 in 2015
* 4 artifacts were problematic in 2016 vs 7 in 2015

The remaining issues are still related to empirical experiments, scalability, customization of experiments in native user environments, and lack of access to a specific hardware (energy counters, machines with > 100 cores). We would also like to discuss how to unify AE appendices and meta in ACM DL

## Special Interest Group on Design Automation (SIGDA)

**Patrick Madden**

Years ago, a survey paper on problematic circuit placement results considered only major publications on "standard" benchmarks, and revealed what looked like a random walk. Out of this grew the ISPD physical design contest.[[112]](#footnote-111)

These contests have been running for more than ten years, and the general approach follow:s

* An industry group defines a problem objective, and produces a set of benchmarks, metrics, and evaluation tools. There are a handful of illustrative benchmark examples.
* Research groups prepare binary versions of their tools, and submit them to the industry group.
* The industry group runs everything on their machines, on a secret set of benchmarks, with the results being reported at ISPD. None of the research groups know the results, or where they stand in the contest, until the announcement.
* After the end of the contest, the secret benchmarks become part of the community reference set.

The contests have had a huge positive impact on design automation research. The industry groups now take academic work seriously, and the problems are pushing the state of the art. Previously, many straw-man benchmarks only served to make authors look good. Having the runs and evaluation performed by an independent group eliminates the temptation of authors to cook the results.

Most of the major SIGDA conferences have contests along these lines now, and design automation papers that try to skirt the widely available benchmarks are unlikely to get accepted. Published results today are much more trustworthy, and there are far fewer groups trying to game the system.

SIGDA has a little bit of influence on the contests, they are largely left up to the conference and contest organizers. To help get industry participation, though, SIGDA routinely gives service awards to the people who put together the benchmarks and metrics. A plaque that someone can show to their boss seems to be an effective motivator. In fact the industry groups now jockey to write the benchmarks and run the contest each year.

The industry groups are usually good at avoiding proprietary data sets, or requiring non-public tools or information. Over the past few years, some of the academic groups have also started releasing source code for their tools.

I very very very much like the SIGMOD "seal-of-reproducibility" -- that seems like a great thing to promote. I would love to see the this applied across ACM.

For being able to repeat experiments -- if there's a need for exotic hardware, that's obviously tough. But if a software tool set can run on a generic Linux box, it seems like it should be possible to set things up so that there's a VM in the digital library. Vint Cerf (and many other folks) have been concerned about bit rot -- perhaps we can promote standard VM environments, so that we only need to make sure we can spin up a VM player, rather than a full system from scratch. We can't make everything reproducible, but for the things where it is feasible -- we should.

IMO, the CS field lags behind other areas for reproducibility. To my knowledge, ACM papers have not been retracted except for plagiarism, and are only rarely amended for errors (someone please correct me if I've missed something). If you look at physics, medicine, chemistry, the communities around them take correctness very seriously, with the prospect of retraction being an axe over everyone’s head. There's an expectation that others should be able to arrive at the same results if they use the same methods, and it's a big deal if that doesn't happen.

As we go down the road of encouraging reproducibility, we'll have to decide what to do if something cannot be verified or reproduced. If, on closer inspection of a paper, there's a logic error, an experimental error, or outright fraud -- what do we do?

## Experimental Mathematical Computation

“Experimental mathematics is the use of a computer to run computations — sometimes no more than trial-and-error tests — to look for patterns, to identify particular numbers and sequences, to gather evidence in support of specific mathematical assertions that may themselves arise by computational means, including search. [9, p. 1][[113]](#footnote-112)

## Simulation and Modelling (SIM)

## Human Computer Interaction (HCI)/Web

**Simon Harper**

My area of HCI/Web seems to be much more linked to “discovery” sciences as opposed to those which are discussed (and which I think of as 'creation' science). In my domain, the collected data is a primary artifact, as well as the algorithms that run over them. In fact our requirements might be the data, the methodologies used to collect that data, and the analysis protocols used in the investigation. We might never use a novel algorithm that needs to be stored or which could automatically execute over the data. Far better the collection protocols to enable replication, and analysis code (maybe in R) to automate the statistics. To me, this feels much more like Open Notebook Science or the ipython notebook[[114]](#footnote-113) inviting the reader to see the work as it progressed as opposed to a set of curated algorithms and corpus they run over. We also need to make sure there is a low overhead in publishing in this way. It may be useful to consider the diversity of the audience, and the flexibility required. I also think the terms are going to need some very clear definitions.

## ACM Special Interest Group on Management of Data (SIGMOD)

## COMM-METRICS

### Publishing measurement data and simulation code used in a variety of Internet measurement papers

**Rich Guerin**

#### Main Publication Venue:

ACM Internet Measurement conference (IMC).[[115]](#footnote-114) Code and data were typically submitted in support of a peer-reviewed paper but without mandating that they be scrutinized to the same level as the paper itself.

#### Goals

* The goal in supplying code and data was primarily to facilitate reproducibility.
* *Allowing* for reproducibility was in general more important than *testing* for reproducibility.
* With respect to the data itself, the goal was to foster greater community access to shared resources.

#### Author Motivation

IMC 2015 will bestow two awards. One award will recognize the outstanding paper at the conference, and all accepted papers are eligible for it. The other award will recognize a paper that contributes a novel dataset to the community. To be eligible for this award, the authors must make their dataset publicly available (e.g., through [DatCat](http://www.datcat.org/) for Internet measurement data or [CRAWDAD](http://crawdad.cs.dartmouth.edu/) for wireless data) by the time of the camera-ready submission.

#### Resources

Resource on ethical sharing of data:

* *The Menlo Report*[[116]](#footnote-115)
* A 2007 paper by Mark Allman and Vern Paxson[[117]](#footnote-116)

Resource on a proposed framework to facilitate reproducibility testing

* Allman & Paxson, “Issues and etiquette concerning use of shared measurement data”[[118]](#footnote-117)

#### Problems

While peer-validation of results obtained using data and/or (simulation) software to the same extent that we verify proofs for theoretical papers would be a very useful thing, it is enormously time-consuming, especially if it involves having to become familiar with a new system.

Some form of "normalization" when it comes to format and platform could help with this aspect, but this approach may only be realistic in very narrow problem domains for which this can be specified.

#### 

#### 

#### General Recommendations

1. Provided that some form of storage persistence is guaranteed, there is little to no benefit in having a common backend repository;
2. The biggest challenges are
   1. offering a logical and easily searchable interface to the available data/software; and
   2. ensuring basic interoperability of formats and their description (this is mostly for data sets, but to some extent for software too).
3. We may need as many different solutions as there are problem spaces, which is not particularly exciting.
4. Possibly have each community define their best practices, with ACM only providing a common front-end for submissions, retrievals, searches, and presentation.

## SIGSOFT/SIGPLAN

## Reproducibility in the Verification Community

**Dirk Beyer**

#### Reproducible Benchmarking and Resource Measurement

Since a couple of years, the verification community has been implementing tools that make it possible to measure resources accurately, during performance benchmarking of CPU-intensive computation tasks. The report below describes the technical requirements for measuring resource consumption and enforcing resource limits, and discusses existing and new technical solutions to achieve the goal.[[119]](#footnote-118)

#### Central Repository of Benchmark Verification Tasks

A requirement for reproducibility of experiments is that the input data for the experiments, here the programs to be verified, are available for the long term and accessible publicly (open source). In the community of automatic verification of software programs, there is a successful approach used to collect benchmarks in a central repository. The repository has existed since 2011, is contiinuously growing, and an inspection of the literature shows that it has had a large impact.

Previously, the experimental evaluations in research publications were based on a handful of verification tasks, because it was considered expensive to construct proper programs / verification tasks for the process. In the last few years, it got established in the literature, that tools for software verification are evaluated on the standard benchmarks from the central repository. For example, the Competition on Software Verification is based on those benchmarks and actively maintains and develops the set.

Repository: <https://github.com/dbeyer/sv-benchmarks/tree/master/c>

Description: <http://sv-comp.sosy-lab.org/2015/benchmarks.php>

#### Yearly Comparative Evaluation of Software Verifiers.

For the research community of software verification, the conference Tools and Algorithms for the Construction and Analysis of Systems (TACAS) has established a yearly comparative evaluation of the research progress by running all available tools of a large benchmark. Organizationally, this is implemented as the "Competition on Software Verification"[[120]](#footnote-119), where the authors have a chance to inspect the results that their verifier obtained, and the results and the verifiers are both archived for reproducing the results at any given time. An overview of the results is published yearly in a report.[[121]](#footnote-120)

#### Support by Conferences and Publishing Vendors

The conferences, or their vendors for proceedings processing, should include

a uniform and simple process for the authors to centrally collect replication packages.

For example, the vendor Conference Publishing Consulting has recently updated

process of preparing the proceedings for the ACM Digital Library, such that collecting Auxiliary Material from the authors for publication in the DL is by default part of the process. Auxiliary Material to be published in the DL can be videos by the authors,

replication packages, supplementary data in a zip archive, or even Virtual Machines.

This is in particular necessary for those papers whose artifacts got positively

evaluated in the Artifact Evaluation process that more and more conferences offer.

## Social Sciences

**Limor Peer**

In social sciences, data has also been a primary artifact , and there are established data archives that have set the standards for data curation for over 50 years. With the recent focus on reproducibility and transparency, additional “research products” -- such as code, protocols and other documentation that enable replication -- are increasingly required. A concern is that these research products are frequently self-deposited in a variety of digital repositories that may not necessarily engage in data curation or other review activities, thus potentially limiting their long term usability. A process we call “data quality review” (DQR) is designed to help journals and repositories, as well as researchers themselves, to validate that best practices have been implemented and that files are ready for long term reuse. The DQR is “a process whereby data and associated files are assessed and required actions are taken to ensure files are independently understandable for informed reuse. This is an active process, involving a review of the files, the documentation, the data, and the code.” There is more on this here <http://www.ijdc.net/index.php/ijdc/article/view/9.1.263/358>.

The concept of data description/documentation is highly developed. See the Data Documentation Initiative (DDI) - <http://www.ddialliance.org/>- the main standard for description in the social sciences. The concept of research objects is gaining ground, and driven primarily by journals that require “replication files” and by reproducibility efforts in fields such as psychology, political science, and economics.

# 

# 

# 

# Appendix B. Project Participants

Simon Adar, Cornell Institute, Code Ocean

Michah Altman, MIT

Dirk Beyer,

Ron Boisvert, NIST

Philippe Bonnet,

### [Peter Buneman,](https://www.inf.ed.ac.uk/people/staff/Peter_Buneman.html) [School of Informatics,The University of Edinburgh](http://www.ed.ac.uk/informatics)

Bruce Childers, Pittsburgh

Jack Davidson, Virginia

Scott Delman, ACM

Nik Dutt, UC-Irvine

Juliana Freire, NYU

Grigori Fursin, Harvard

Adam G, ACM

Wayne Graves, ACM

Mark Gross, Data Conversion Lab

David Grove, IBM Research

Roch Guerin, Washington U., St. Louis

Simon Harper, U. Manchester

Michael Heroux

T.R. Hopkins, University of Kent

Stratos Idreos

Joe Konstan

Michael Lesk, Rutgers University

Randall J. LeVeque, Applied Mathematics, University of Washington, Seattle

Patrick Madden, Binghamton U

Mike Miksis, SIAM

Sheila Morrissey, Ithaca

Limor Peer, Yale

Wilfred Pinfold

Craig Rodkin, ACM

Pat Ryan, ACM

Bernard Rous, ACM

Henning Schulzrinne, Columbia University, Networking

Gianluca Setti, IEEE

Gwendal Simon, Telecom Bretagne

Stephen Spencer

Michela Taufer, Delaware

Lin U

Alex Wade, Microsoft Research

1. <https://www.acm.org/publications/digital-library>, accessed 21 December 2015. [↑](#footnote-ref-0)
2. Marcus, E., “Reproducibility, rigor and reliability, robustness: everybody’s talking, but what are we talking about,” (presentation) *STM Innovations Seminar,* Dec 2015, <https://www.youtube.com/watch?v=MagtNQjeE_U&index=4&list=PLv9XTce9yAu4KeP8GB2VWegnRoxecBTeW>, accessed 2 Feb 2016. [↑](#footnote-ref-1)
3. Begley, C. G. & Lee M. Ellis, “Drug development: Raise standards for preclinical cancer research, Nature 483, 531–533 (29 March 2012), <http://dx.doi.org/10.1038/483531a>. [↑](#footnote-ref-2)
4. Dienes, Z., “The reproducibility crisis in science and how to solve it,” (presentation) *STM Innovations Seminar,* 2 Dec 2015, <https://www.youtube.com/watch?list=PLv9XTce9yAu4KeP8GB2VWegnRoxecBTeW&v=GvHUj4VCllI>, accessed 29 Jan 2016. [↑](#footnote-ref-3)
5. Baker, D., “The Reinhart-Rogoff debt-to-GDP error: Why it matters,” *CEPR Blog,* 18 April 2013, <http://cepr.net/blogs/cepr-blog/the-reinhart-rogoff-debt-to-gdp-error-why-it-matte>, accessed 21 Dec 2015. [↑](#footnote-ref-4)
6. Shanahan, D., “[Why perpetuate a 300-year-old anachronism? Reincarnating the research article into a ‘living document’](http://blogs.lse.ac.uk/impactofsocialsciences/2015/05/13/reincarnating-the-research-article-into-a-living-document/),” *The Impact Blog,* London School of Economics and Political Science, 13 May 2015, <http://blogs.lse.ac.uk/impactofsocialsciences/2015/05/13/reincarnating-the-research-article-into-a-living-document/> [↑](#footnote-ref-5)
7. "’Play it again Sam’ - Research reproducibility and reputation management,” *STM Innovations Seminar,* Dec 2015, <http://www.stm-assoc.org/events/innovations-seminar-2015/>, accessed 27 Jan 2016. [↑](#footnote-ref-6)
8. 2013 NFAIS/NISO Recommended Practices for Online Supplemental Journal Articles Materials, <http://www.niso.org/apps/group_public/download.php/10055/RP-15-2013_Supplemental_Materials.pdf>, p. v., accessed 4 Feb 2016. [↑](#footnote-ref-7)
9. *RCUK common principles on data policy,* revised July 2015, <http://www.rcuk.ac.uk/research/datapolicy/>, accessed 15 Jan 2016. [↑](#footnote-ref-8)
10. “University of Manchester Research Data Management Policy,” *University of Manchester Libraries Web Site,* <http://www.library.manchester.ac.uk/services-and-support/staff/research/services/research-data-management/policy/>, accessed15 Jan 2016. [↑](#footnote-ref-9)
11. “Data management planning,” *University of Manchester Libraries Web Site,* <http://www.library.manchester.ac.uk/services-and-support/staff/research/services/research-data-management/data-management-planning/>, accessed 15 Jan 2016. [↑](#footnote-ref-10)
12. NISO, op. cit., p. 4. [↑](#footnote-ref-11)
13. “Science as an open enterprise: The Royal Society Science Policy Centre report 02/12,” June 2012, *The Royal Society,* <https://royalsociety.org/~/media/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf>, accessed 21 Jan 2016. [↑](#footnote-ref-12)
14. [Roger D. Peng](http://www.ncbi.nlm.nih.gov/pubmed/?term=Peng%20RD%5Bauth%5D), “Reproducible research in computational science, [*Science,* 2011 Dec 2; 334(6060): 1226–1227](http://www.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&retmode=ref&cmd=prlinks&id=22144613), http://dx.doi.org/[10.1126/science.1213847](http://dx.doi.org/10.1126%2Fscience.1213847) [↑](#footnote-ref-13)
15. Ibid. [↑](#footnote-ref-14)
16. V. Stodden, et al, “Setting the default to reproducible,” *Reproducibility in Computational and Experimental Mathematics*, 2013, <https://icerm.brown.edu/tw12-5-rcem/icerm_report.pdf>, accessed 4 Feb 2016. [↑](#footnote-ref-15)
17. Statistics -- Vocabulary and symbols -- Part 2: Applied statistics, ISO 3534-2:2006, <http://www.iso.org/iso/catalogue_detail.htm?csnumber=40147>, 5.2.5, accessed 4 Feb 2016. [↑](#footnote-ref-16)
18. Ibid, 3.3.5, 3.3.6 [↑](#footnote-ref-17)
19. *JCGM 200:2012, International vocabulary of metrology – Basic and general concepts and associated terms (VIM)*, 3rd edition, 2008 version with minor corrections [↑](#footnote-ref-18)
20. Stodden et al, p. 2. [↑](#footnote-ref-19)
21. ISO 3534-2:2006, 3.3.10, 3.3.11 [↑](#footnote-ref-20)
22. JCGM, op. cit. [↑](#footnote-ref-21)
23. Marcus, E., op. cit. [↑](#footnote-ref-22)
24. Dienes, op cit. [↑](#footnote-ref-23)
25. Stodden et al., p. 1. [↑](#footnote-ref-24)
26. Stodden et al., p 15. [↑](#footnote-ref-25)
27. Nosek, B. et al, “Promoting an open research culture,” *Science*, 26 Jun 2015, Vol. 348, Issue 6242, pp. 1422-1425, <http://dx.doi.org/10.1126/science.aab2374>. [↑](#footnote-ref-26)
28. “Science Code Manifesto,” *Climate Code Foundation,* 2013, <http://sciencecodemanifesto.org/>, accessed 21 Jan 2016. [↑](#footnote-ref-27)
29. *Clarifications of EPSRC expectations on research data management,* 9 Oct 2014, <https://www.epsrc.ac.uk/files/aboutus/standards/clarificationsofexpectationsresearchdatamanagement/>, accessed 4 Feb 2016. [↑](#footnote-ref-28)
30. Stodden et al, p. 3 [↑](#footnote-ref-29)
31. “Why should I be part of this,” and “How much overhead is it?”, *ACM SIGMOD 2015 Reproducibility,* <http://db-reproducibility.seas.harvard.edu/>, accessed 5 Jan 2016. [↑](#footnote-ref-30)
32. Manolescu, I and Manegold, S., “Performance evaluation in database research: Principles and experiences,” *ICDE 2008,* Slide 235, <http://pages.saclay.inria.fr/ioana.manolescu/SLIDES/ManolescuManegoldICDE2008.pdf>, accessed 5 Jan 2016. [↑](#footnote-ref-31)
33. Shaklee, P. and Helena Cousijn, “Can data be peer-reviewed?” Elsevier Connect, 6 July 2015, <https://www.elsevier.com/connect/can-data-be-peer-reviewed>, accessed 21 Jan 2016. [↑](#footnote-ref-32)
34. “A list of data Journals,” *Trac website,*  <http://proj.badc.rl.ac.uk/preparde/blog/DataJournalsList>, accessed 21 Jan 2016. [↑](#footnote-ref-33)
35. Dienes, op. cit. [↑](#footnote-ref-34)
36. Stodden et al, p. 6. [↑](#footnote-ref-35)
37. Manolescu, I and Manegold, S., “Performance Evaluation in Database Research: Principles and Experiences,” *ICDE 2008,* Slide 231, <http://pages.saclay.inria.fr/ioana.manolescu/SLIDES/ManolescuManegoldICDE2008.pdf>, accessed January 5, 2015 [↑](#footnote-ref-36)
38. Dienes, op cit. [↑](#footnote-ref-37)
39. Stodden et al, “The role of employers and research managers,” p. 7 [↑](#footnote-ref-38)
40. Stodden et al., “Teaching reproducibility skills,” p. 7 [↑](#footnote-ref-39)
41. Stodden et. al, “Appendix G The teaching and training of reproducibility skills,” p. 17. [↑](#footnote-ref-40)
42. Stodden et al. “Appendix E Tools to aid in reproducible research,” p. 14 [↑](#footnote-ref-41)
43. Stodden et al, “The Role of Funding Agency Policy,” p. 5-6 [↑](#footnote-ref-42)
44. DIenes, op cit. [↑](#footnote-ref-43)
45. Lane, S., “All trials registered, all results reported--the publisher’s role,” *STM Innovations Seminar,* Dec 2015, <https://www.youtube.com/watch?v=5KnbDaVSx20&list=PLv9XTce9yAu4KeP8GB2VWegnRoxecBTeW&index=2>, accessed 2 Feb 2016.

    [↑](#footnote-ref-44)
46. Stodden et al, “The Rrle of journal policy” p. 6. [↑](#footnote-ref-45)
47. Stodden et al, p. 6. [↑](#footnote-ref-46)
48. Stodden et al, “Appendix D Best Practices for Publishing Research,” p. 13-14. [↑](#footnote-ref-47)
49. “Briefing paper for RDA Publishing Data Interest Group,” RDA Sixth Plenary Meeting Programme, <https://rd-alliance.org/bof-journal-research-data-policies-p6-bof-session.html>, accessed 25 Jan 2016. [↑](#footnote-ref-48)
50. “Best practices,” *Research Data Management Service Group,* Cornell University web site, <http://data.research.cornell.edu/content/best-practices>, accessed 21 Jan 2016. [↑](#footnote-ref-49)
51. Altman, M. et al, “An Introduction to the Joint Principles for Data Citation,” *Bulletin of the Association for Information Science and Technology*, February/March 2015, Vol. 41, Number 3, [https://www.asis.org/Bulletin/Feb-15/FebMar15\_RDAP\_Altman\_EtAl.pdf.](https://www.asis.org/Bulletin/Feb-15/FebMar15_RDAP_Altman_EtAl.pdf) accessed 21 Jan 2016. [↑](#footnote-ref-50)
52. “Data Citation,” Cornell RDMSG, <http://data.research.cornell.edu/content/data-citation>, accessed 21 Jan 2016. [↑](#footnote-ref-51)
53. Diepenbroek, M., “Data Literature Interlinking (DLI): A universal service,” *STM Innovations Seminar,* December 2016, <https://www.youtube.com/watch?v=GFs7-7H1xck&index=7&list=PLv9XTce9yAu4KeP8GB2VWegnRoxecBTeW>, accessed 2 Feb 2016. [↑](#footnote-ref-52)
54. “File formats,” Cornell University RDMSG, <http://data.research.cornell.edu/content/file-formats>, accessed 21 Jan 2016. [↑](#footnote-ref-53)
55. Peer et al, “Committing to data quality review,” *International Journal of Digital Curation,* 2014, 19(1): 263-291. <http://dx.doi.org/10.2218/ijdc.v9i1.317>. [↑](#footnote-ref-54)
56. Shaklee, op cit. [↑](#footnote-ref-55)
57. “WG RDA/WDS Publishing Data Workflows - P6 meeting session,” *RDA Sixth Plenary Programme,* <https://rd-alliance.org/wg-rdawds-publishing-data-workflows-p6-meeting-session.html>, accessed 25 Jan 2016. [↑](#footnote-ref-56)
58. “Data publishing workflows (slide presentation), *RDA Sixth Plenary Programme,* <https://rd-alliance.org/sites/default/files/attachment/01-WG-publishing-data-workflows.pdf>, accessed 25 Jan 2016. [↑](#footnote-ref-57)
59. “ACS publications partners with Digital Science’s Figshare to promote open data discovery and use,” (news release) *American Chemical Society Website,* 5 Nov 2015, [http://www.acs.org/content/acs/en/pressroom/newsreleases/2015/november/acs-publications-partners-with-digital-sciences-figshare-to-promote-open-data-discovery-and-use.htm](http://www.acs.org/content/acs/en/pressroom/newsreleases/2015/november/acs-publications-partners-with-digital-sciences-figshare-to-promote-open-data-discovery-and-use.html), accessed 3 Feb 2016. [↑](#footnote-ref-58)
60. “Recommended data repositories,” *Scientific Data,* <http://www.nature.com/sdata/data-policies/repositories>, accessed 2 Feb 2016. [↑](#footnote-ref-59)
61. <http://www.re3data.org/>, accessed 2 Feb 2016. [↑](#footnote-ref-60)
62. DLI Service, <http://dliservice.research-infrastructures.eu/#/>, accessed 2 Feb 2016. [↑](#footnote-ref-61)
63. Diepenbroek, op cit. [↑](#footnote-ref-62)
64. Data Citation Synthesis Group, Martone M. (ed.), “Joint Declaration of Data Citation Principles.” San Diego CA: *FORCE11;* 2014, <https://www.force11.org/datacitation>, accessed Jan. 21, 2016. [↑](#footnote-ref-63)
65. Rauber, A. et al, “Data citation of evolving data recommendations of the Working Group on Data Citation (WGDC),” Revision 24 September 2015, <https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_150924.pdf>, accessed 25 Jan 2016. [↑](#footnote-ref-64)
66. RMap Project, <http://rmap-project.info/rmap/>, accessed 21 Jan 2016 [↑](#footnote-ref-65)
67. Webster, K. and Cochrane, E. “Software curation as a digital preservation service,” <https://vimeo.com/channels/cni/129809572>, *Coalition for Networked Information (CNI) Spring 2015 Membership Meeting,* April 13-14, 2015, accessed 14 Jan 2016. [↑](#footnote-ref-66)
68. bwFLA project from University of Freiburg, <http://bw-fla.uni-freiburg.de/>, accessed 14 Jan 2016. [↑](#footnote-ref-67)
69. Stodden et al, “Appendix F Copyright and licensing”, p 16-17. [↑](#footnote-ref-68)
70. Matthews, B. et al., “Appendix A: An OWL ontology for significant properties of software,” *The Significant Properties of Software: A Study,* 2008, <http://www.jisc.ac.uk/media/documents/programmes/preservation/significantpropertiesofsoftware-final.doc>, accessed 18 Jan 2016. [↑](#footnote-ref-69)
71. Software ontology, from Software Ontology for Resource Description (SWORD) European Informatics Institute + University of Manchester, <http://theswo.sourceforge.net/>, accessed 18 Jan 2016 [↑](#footnote-ref-70)
72. <http://www.keep-totem.co.uk/>, used by Dioscuri (IBM, National Library of the Netherlands, PLANETS project), BwFLA EU), accessed 18 Jan 2016. [↑](#footnote-ref-71)
73. [www.nsrl.nist.gov](http://www.nsrl.nist.gov), accessed 18 Jan 2016. [↑](#footnote-ref-72)
74. McDonough, J., “Some assembly required: Reflections on XML semantics, digital preservation and the construction of knowledge,” In *Proceedings of Balisage: The Markup Conference 2013*, August 2013, Balisage Series on Markup Technologies, vol. 10 (2013), <http://dx.doi.org/10.4242/BalisageVol10.McDonough01>. [↑](#footnote-ref-73)
75. Jones, C. et. al., “Software reuse, repurposing and reproducibility,” iPRES 15, November 2015. [↑](#footnote-ref-74)
76. Rosenthal, D. S. H., “Emulation & virtualization as preservation strategies,” DSHR Blog, November 2015, [↑](#footnote-ref-75)
77. Contaxis, N. “Preserving In-House Developed Software,” (poster) iPRES 15, November 2015. [↑](#footnote-ref-76)
78. “Guidelines,” ACM SIGMOD 2015 Reproducibility, <http://db-reproducibility.seas.harvard.edu/#Guidelines>, accessed January 5, 2016. [↑](#footnote-ref-77)
79. Manolescu, I and Manegold, S., “Performance evaluation in database research: Principles and experiences,” *ICDE 2008,* Slides 160, 179, <http://pages.saclay.inria.fr/ioana.manolescu/SLIDES/ManolescuManegoldICDE2008.pdf>, accessed January 5, 2015 [↑](#footnote-ref-78)
80. “Guidelines,” ACM SIGMOD 2015 Reproducibility, <http://db-reproducibility.seas.harvard.edu/#Guidelines>, accessed January 5, 2016.. [↑](#footnote-ref-79)
81. Peer, et al, “Committing to data quality review,” *International Journal of Digital Curation,* 2014. 19(1): 263-291. <http://dx.doi.org/10.2218/ijdc.v9i1.317>. [↑](#footnote-ref-80)
82. Thain, D. et. al., “Techniques for preserving scientific software executions: Preserve the mess or encourage cleanliness?” (presentation), 12th International Conference on Digital Preservation (iPRES) 2015, November 2015. [↑](#footnote-ref-81)
83. Manolescu, I and Manegold, S., “Performance evaluation in database research: Principles and experiences,” *ICDE 2008,* Slides 152-153, <http://pages.saclay.inria.fr/ioana.manolescu/SLIDES/ManolescuManegoldICDE2008.pdf>, accessed January 5, 2015 [↑](#footnote-ref-82)
84. Thain, D. et. al., “Techniques for preserving scientific software executions: Preserve the mess or encourage cleanliness?” (presentation), *12th International Conference on Digital Preservation (iPRES) 2015*, November 2015. [↑](#footnote-ref-83)
85. Stodden et al, “Appendix F Copyright and licensing”, p 17. [↑](#footnote-ref-84)
86. “Policy Statement on Software,” American Astronomical Society Website, <http://journals.aas.org/policy/software.html>, accessed 4 Feb 2016. [↑](#footnote-ref-85)
87. Marcus, E., op cit. [↑](#footnote-ref-86)
88. *Reporting checklists for life science journals,* <http://www.nature.com/authors/policies/checklist.pdf>, accessed 2 Feb 2016. [↑](#footnote-ref-87)
89. Hufton, A. L., “Promoting transparent research methods, protocols and data to reduce irreproducibility,” (presentation) *STM Innovations Seminar,* December 2015, <http://www.stm-assoc.org/2015_12_02_Innovations_Seminar_Hufton_Publishing_Innovations.pdf>, accessed 4 Feb 2016. [↑](#footnote-ref-88)
90. \* <https://hal.inria.fr/hal-01054763>

    \* [http://cTuning.org/papers/CPC2015.pdf](http://ctuning.org/papers/CPC2015.pdf)

    \* <http://dl.acm.org/citation.cfm?id=2618142>

    \* <http://cknowledge.org/repo>

    \* <http://c-mind.org/browse>

    \* <http://ctuning.org/reproducibility> [↑](#footnote-ref-89)
91. Thomas J. Leeper, “What's in a name? The concepts and language of replication and reproducibility,” *Open Science Language,* May 2015, <http://thomasleeper.com/2015/05/open-science-language/> , accessed January 11, 2016 [↑](#footnote-ref-90)
92. http://www.researchobject.org/ [↑](#footnote-ref-91)
93. https://www.docker.com/ [↑](#footnote-ref-92)
94. http://ipython.org/notebook.html [↑](#footnote-ref-93)
95. http://fair-dom.org/ [↑](#footnote-ref-94)
96. http://www.mygrid.org.uk/ [↑](#footnote-ref-95)
97. http://www.seek4science.org/ [↑](#footnote-ref-96)
98. [10.1145/355637.355643](http://dx.doi.org/10.1145/355637.355643) [↑](#footnote-ref-97)
99. [10.1145/2699464](http://dx.doi.org/10.1145/2699464) . [↑](#footnote-ref-98)
100. “Submisssions,” *IEEE/ACM Transactions on Networking* web site, <http://www.ifp.illinois.edu/ton/submissions.html#publication>, accessed 21 December 2015. [↑](#footnote-ref-99)
101. Setti, G., “Reproducibility: what?, why?, when? Experiences within the IEEE and ideas to move forward,” (presentation) *STM Innovations Seminar,* Dec 2015, <https://www.youtube.com/watch?v=bZAWmV8MX-c&index=3&list=PLv9XTce9yAu4KeP8GB2VWegnRoxecBTeW>, accessed 2 Feb 2016. [↑](#footnote-ref-100)
102. Bailey, D. H., “Twelve ways to fool the masses when giving performance results on parallel computers,” S*upercomputing Review,* June 11, 1991, Ref: Aug. 1991, pg. 54--55. [↑](#footnote-ref-101)
103. Southern, G. and Renau, J., “Deconstructing PARSEC scalability,” <https://users.soe.ucsc.edu/~renau/docs/wddd15.pdf>, accessed 20 January 2016. [↑](#footnote-ref-102)
104. Call for Papers, Workshop on Reproducibility, November 2015. [↑](#footnote-ref-103)
105. [10.1145/513001.513005](http://dx.doi.org/10.1145/513001.513005) [↑](#footnote-ref-104)
106. Heroux, M., “Editorial: ACM TOMS Replicated Computational Results initiative,” *ACM Trans. Math. Softw.* 41, 3, Article 13 (June 2015)*,* <http://dx.doi.org/10.1145/2743015>. [↑](#footnote-ref-105)
107. Van Zee, F. G. and van de Geijn, R. A.. “BLIS: A framework for rapidly instantiating BLAS functionality,” *ACM Trans. Math. Softw.* 41, 3, Article 14 (June 2015), <http://dx.doi.org/10.1145/2764454>. [↑](#footnote-ref-106)
108. Willenbring, J. M. “Replicated Computational Results (RCR) report for “BLIS: A framework for rapidly instantiating BLAS functionality”. ACM Trans. Math. Softw. 41, 3, Article 15 (June 2015), 4 pages, <http://dx.doi.org/10.1145/2738033>. [↑](#footnote-ref-107)
109. <http://ctuning.org/event/ae-cgo2015/> [↑](#footnote-ref-108)
110. Fursin, G. and Dubach, C., “Community-driven reviewing and validation of publications, In *Proceedings of the 1st ACM SIGPLAN Workshop on Reproducible Research Methodologies and New Publication Models in Computer Engineering (TRUST '14)*. ACM, New York, NY, Article 5, <http://dx.doi.org/10.1145/2618137.2618142> [↑](#footnote-ref-109)
111. <http://www.nsf.gov/awardsearch/showAward?AWD_ID=1535232>) [↑](#footnote-ref-110)
112. ISPD 2015 Blockage-Aware Detailed Routing-Driven Placement Contest, <http://www.ispd.cc/contests/15/ispd2015contest.html> [↑](#footnote-ref-111)
113. Bailey, D. et. al., “Opportunities and challenges in 21st century experimental mathematical computation: ICERM Workshop Report,” August 28, 2014, <http://www.davidhbailey.com/dhbpapers/ICERM-2014.pdf>, accessed 20 Jan 2016. [↑](#footnote-ref-112)
114. <http://ipython.org/notebook.html> [↑](#footnote-ref-113)
115. <http://conferences.sigcomm.org/imc/2015/cfp.html> [↑](#footnote-ref-114)
116. *The Menlo Report Ethical Principles Guiding Information and Communication Technology Research,* August 2012, <http://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/menlo_report_actual_formatted.pdf>, accessed 4 Feb 2016. [↑](#footnote-ref-115)
117. Allman, M. and Vern Paxson, “Issues and etiquette concerning use of shared measurement data,” In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement* (IMC '07). ACM, New York, NY, USA, 135-140, <http://doi.acm.org/10.1145/1298306.1298327>. [↑](#footnote-ref-116)
118. Tazaki, H. et al., “Direct code execution: revisiting library OS architecture for reproducible network experiments,” In *Proceedings of the ninth ACM conference on Emerging networking experiments and technologies* (CoNEXT '13). ACM, New York, NY, USA, 217-228, <http://doi.acm.org/10.1145/2535372.2535374>. [↑](#footnote-ref-117)
119. <http://www.sosy-lab.org/~dbeyer/Publications/2015-SPIN.Benchmarking_and_Resource_Measurement.pdf> [↑](#footnote-ref-118)
120. <http://sv-comp.sosy-lab.org/2015/> [↑](#footnote-ref-119)
121. <http://www.sosy-lab.org/~dbeyer/Publications/2015-TACAS.Software_Verification_and_Verifiable_Witnesses.pdf> [↑](#footnote-ref-120)