Beyond Open Data
A Model for Linking Digital Artifacts to Enable Reproducibility of Scientific Claims

Victoria Stodden
School of Information Sciences
University of Illinois Urbana-Champaign
vcs@stodden.net

P-RECS'20
3rd International Workshop on Practical Reproducible Evaluation of Systems
June 24, 2020

Agenda

1. Framing: Open Data and Open Science
2. Gaps: Reproducibility?
3. Proposal: A Linked Model to Enable Reproducibility

Open Data and Open Science

Many efforts toward Open Data and Open Science:

- FORCE11
- FAIR data
- Science.gov
- NYC Open Data
- Journal data availability policies / new data oriented publications
- Funding agency efforts e.g. Wellcome Trust, NSF Data Management Plan
- Legislative efforts around open data in the US, EU..
- many many more..

A principled approach: *artifact availability for the verification of claims*

NASEM Report Definitions

Reproducibility is obtaining ***consistent results using the same input data, computational steps, methods****, and code, and conditions of analysis*. This definition is synonymous with “computational reproducibility”

Replicability is obtaining ***consistent results across studies aimed at answering the same scientific question***, each of which has obtained its own data. Two studies may be considered to have replicated if they obtain consistent results given the level of uncertainty inherent in the system under study.

National Academies of Sciences, Engineering, and Medicine. 2019. Reproducibility and Replicability in Science. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303.

Computational Reproducibility

Traditionally two branches to the scientific method:

- Branch 1 (deductive): mathematics, formal logic.
- Branch 2 (empirical): statistical analysis of controlled experiments.

Now, new branches due to technological changes?

- Branch 3,4? (computational): large scale simulations / data driven computational science.

The Ubiquity of Error

The central motivation for the scientific method is to root out error:

- Deductive branch: the well-defined concept of the proof,
- Empirical branch: the machinery of hypothesis testing, appropriate statistical methods, structured communication of methods and protocols.

Claim: Computation and Data Science present only *potential* third/fourth branches of the scientific method, until the development of comparable standards.

Build on Open Data success

- Recognize the value of data is in the information it contains
- Extracting that information is done in software

- We need:
  - Complete soup-to-nuts computational discovery pipelines enable the verification of claims
  - Link: data, with code, to claims in the scholarly record for reproducibility
  - Publish these linked artifacts with the claim
  - Artifacts can live in different repositories when associated by persistent links

Claim 1: Policies Must Recognize the Relationships Between Digital Artifacts That Generate Knowledge

Example: Data Management Plans

DMPs tend not to explicitly mention software or connections to published results that depend on the data

Proposal: Revise DMPs to focus on computational pipelines that support claims in the scholarly record, as well as individual data (e.g. if no claims have yet been made)

Claim 2: Policies Often Prioritize Open Data, and Open Code Lags

- Journals are increasingly requiring the publication of data with journal articles, however policies regarding code or workflow publication lag, sometimes on the order of years (or never).

- Policies could rely on Crossref to ensure artifacts that support published claims are persistently linked.

Challenge 1: Data and Software Can Change Frequently

- Datasets are subject to change and revision, often with high velocity, and such changes may require new unique identifiers for new versions of datasets that are derived from previous versions.

- Software can be updated and changed (often for different reasons and in different ways than data) and unique identifiers may also be appropriate to assign to new versions.

Challenge 2: Data and Software Ownership and Rights Structures Are Different

Data and code may have different authors or creators than the manuscript, and many (different) contributors at different points in time.

Data is different to software in ways that alter the rights assigns and ability to publish. Data may be obtained from many sources each with its own terms of use.

Software, like text and figures, is subject to copyright in the United States, whereas data generally are not (although can be in other countries).

Software may contain proprietary and/or potentially patentable algorithms.

⇨ Software and non-data artifacts cannot be treated like data, from a policy perspective.

Challenge 3: Data and Software Engender Different Preservation Strategies

Software, by definition, is meant to execute on a computational system.

Data, by contrast, can exist in a fixed form, without adapting to a particular computational system.

Software requires specialized maintenance and has different types of dependences to data since it relies on other software components to execute.

Data can reach a size and scale that software is very unlikely to reach.

⇨ Software and non-data artifacts cannot be treated like data, from an archiving perspective.

Challenge 4: Ethical Considerations Differ for Data and Code

Policy frameworks are developing to enable artifact sharing and access.

Data may contain personally identifiable information e.g. human subjects research, and federal rules may exist controlling its release, HIPAA, FERPA.

Software does not have release restrictions due to potential privacy violations although policies are evolving.

Software is subject to copyright and potentially patents.

⇨ Software and non-data artifacts cannot be treated like data, from a policy perspective.

A linked approach to artifact publication is necessary to support reproducibility

Three identifiers: the published article containing the scholarly claims; the data supporting those claims; and the software that analyzed the data to discover or generate the claims, rely on each other to support research reproducibility.

The metadata schemas associated with DOIs (e.g. https://schema.datacite.org/) contain relational information regarding the interconnectedness of digital scholarly objects, implying that such a change in information representation requires very little new infrastructure to recognize the inseparability of data and software, and the claims they support.

A community effort, extending that under way for Open Data, is essential.

Community Approach

Researchers
(processes)

Funders
(policy)

Universities/ institutions
(hiring/promotion; programmatic change)

Universities/ libraries
(empowering w/tools)

Publishers
(TOP guidelines)

Scientific Societies

Regulatory Bodies
(OSTP)

Conclusion

Two reinforcing research trends are converging:

> Scientific projects will become massively more compute and data intensive,
> Research computing will become dramatically more transparent.

These can admit a computable scholarly record, if care is take to connect the artifacts produced.

An approach designed to allow a consumer of published research to access a collection of digital artifacts, including open data and evolving away from a sole focus on open data, will permit computational reproducibility, including transparency in how the claims were derived.