# Enabling Access to Research Data: Key Steps Forward

**Victoria Stodden**
School of Information Sciences
University of Illinois at Urbana-Champaign

**"Towards New Principles for Enhanced Access to Public Data for Science, Technology and Innovation"**
OECD Workshop, Paris
March 13, 2018

# Two Framing Ideas

- Include access to **code** together with data, to allow computational reproducibility of scientific results.

- Consider the entire '**research ecosystem**' including researchers, digital scholarly object repository mangers and librarians, funding agencies, journal editors, and other stakeholders when addressing openness issues.

# Computational Reproducibility

***An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete ... set of instructions [and data] which generated the figures.***

David Donoho, 1998 http://statweb.stanford.edu/~wavelab/Wavelab_850/wavelab.pdf

REPRODUCIBILITY

# Enhancing reproducibility for computational methods

## Data, code, and workflows should be available and cited

*By* **Victoria Stodden,[1] Marcia McNutt,[2] David H. Bailey,[3] Ewa Deelman,[4] Yolanda Gil,[4] Brooks Hanson,[5] Michael A. Heroux,[6] John P.A. Ioannidis,[7] Michela Taufer[8]**

Over the past two decades, computational methods have radically changed the ability of researchers from all areas of scholarship to process and analyze data and to simulate complex systems. But with these advances come challenges that are contributing to broader concerns over irreproducibility in the scholarly literature, among them the lack of transpar-

Sufficient metadata should be provided for someone in the field to use the shared digital scholarly objects without resorting to contacting the original authors (i.e., http://

**Access to the computational steps taken to process data and generate findings is as important as access to data themselves.**

Stodden, Victoria, et al. "Enhancing reproducibility for computational methods." *Science* 354(6317) (2016)

ness Promotion (TOP) guidelines (*1*) and recommendations for field data (*2*), emerged from workshop discussions among funding agencies, publishers and journal editors, industry participants, and researchers repre-

to understanding how computational results were derived and to reconciling any differences that might arise between independent replications (*4*). We thus focus on the ability to rerun the same computational steps on the same data the original authors used as a minimum dissemination standard (*5*, *6*), which includes workflow information that explains what raw data and intermediate results are input to which computations (*7*). Access to the data and code that underlie discoveries can also enable downstream scientific contributions, such as meta-analyses, reuse, and other efforts that include

results are the data, the computational steps that produced the findings, and the workflow describing how to generate the results using the data and code, including parameter settings, random number seeds, make files, or

All data, code, and workflows, including software written by the authors, should be cited in the references section (*10*). We suggest that software citation include software version information and its unique identifier in addi-

# Reproducibility Enhancement Principles

1: To facilitate reproducibility, **share the data, software, workflows**, and details of the computational environment in open repositories.

2: To enable discoverability, **persistent links** should appear in the published article and include a permanent identifier for data, code, and digital artifacts upon which the results depend.

3: To enable credit for shared digital scholarly objects, **citation** should be standard practice.

4: To facilitate reuse, adequately **document** digital scholarly artifacts.

5: Journals should conduct a **Reproducibility Check** as part of the publication process and enact the TOP Standards at level 2 or 3.

6: Use **Open Licensing** when publishing digital scholarly objects.

7: Funding agencies should instigate **new research** programs and pilot studies.

# Querying the Scholarly Record

- Show a table of effect sizes and p-values in all phase-3 clinical trials for Melanoma published after 1994;

- Name all of the image denoising algorithms ever used to remove white noise from the famous “Barbara” image, with citations;

- List all of the classifiers applied to the famous acute lymphoblastic leukemia dataset, along with their type-1 and type-2 error rates;

- Create a unified dataset containing all published whole-genome sequences identified with mutation in the gene BRCA1;

- Randomly reassign treatment and control labels to cases in published clinical trial X and calculate effect size. Repeat many times and create a histogram of the effect sizes. Perform this for every clinical trial published in the 2003 and list the trial name and histogram side by side.

Donoho & Gavish, “Three Dream Applications of Verifiable Computational Results,” CiSE, 2012