Open Data & Reproducibility

Love Data Week

February 15, 2018

Danielle Kane

Data Management & Curation Librarian

What are “data”

Definitions vary from discipline to discipline

- Scientific data is defined as information collected using specific methods for a specific purpose of studying or analyzing.

- Evidence which is used or created to generate new knowledge and interpretations. https://kaptur.wordpress.com/2013/01/23/what-is-visual-arts-research-data-revisited/

Terminology

Open Access

The free, immediate, online availability of research articles coupled with the rights to use these articles fully in the digital environment.

Open Science

The movement to make scientific research, data and dissemination accessible to all levels of an inquiring society, amateur or professional.

Open Data

Part of the Open Access (OA) movement. Data that can be freely used, re-used, re-distributed (under CC-BY license)

Or Open Research Object

"Research Object" gaining popularity

Types of Research Objects

- Data (numeric, written, audiovisual…)
- Software code
- Workflows and methodologies
- Slides, logs, lab books, sketchbooks, notebooks, etc.

What makes data “open”?

- Open data is data that can be freely used, re-used and redistributed by anyone
  - Subject only, at most, to the requirement to attribute and sharealike.
  - Availability and Access
  - Re-use and Redistribution
  - Universal Participation

Read more@ https://sparcopen.org/open-data/

Interoperability

- Diverse systems and organizations working together (inter-operate).
- The ability to componentize and to ‘plug together’ components which is essential to building large, complex systems.

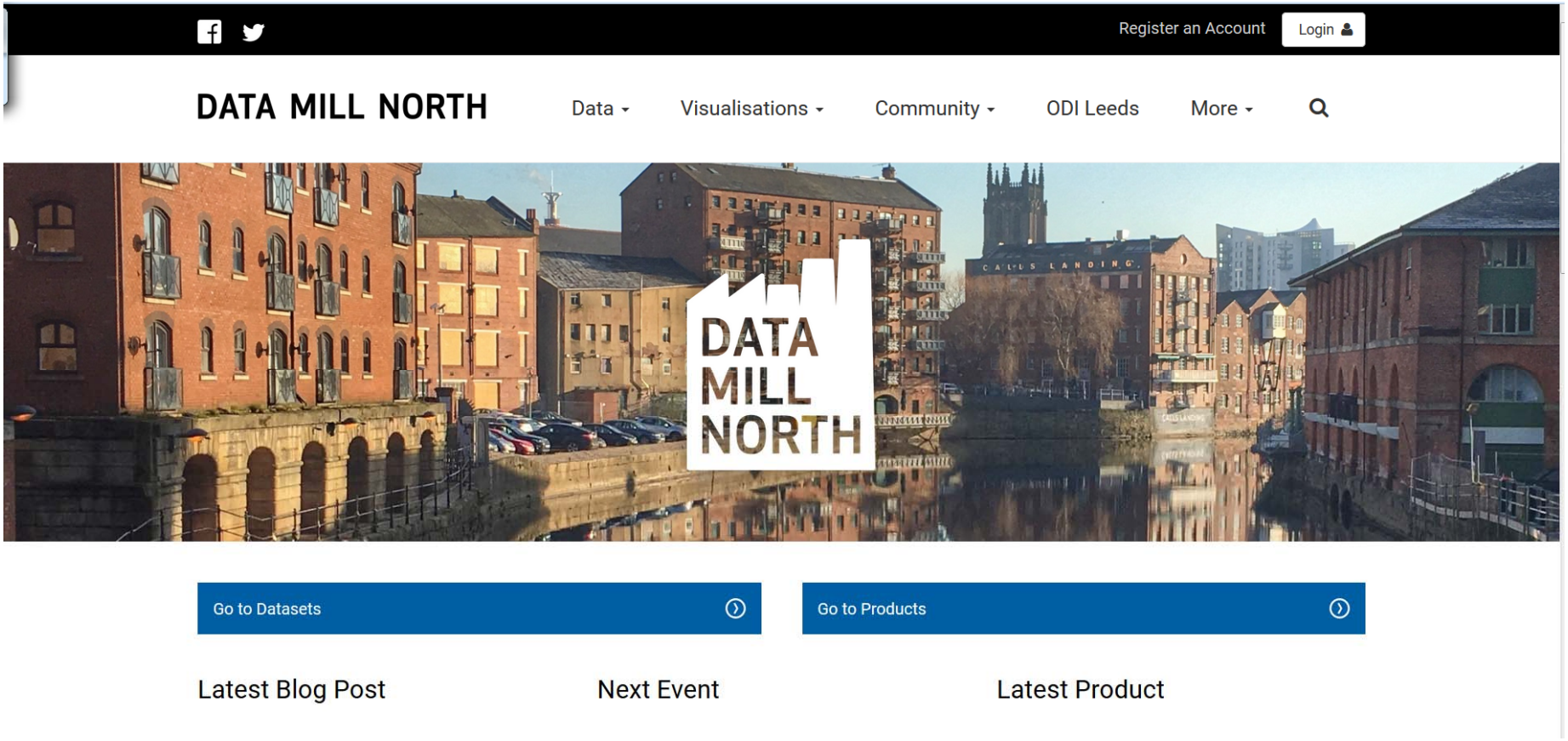
Open Data in Research

- Openness in research is about greater transparency, accessibility and accountability
  - Open Access (OA) came out of High Energy Physics research community
  - Strong government/funder support
- Lower barriers to accessing the outputs of publicly funded research
- Speed up the research process
- Strengthen the quality, integrity and longevity of the scholarly record

Why Open Data?

- Transparency
- Participation
- Self-empowerment
- Improve or create new private products and services
- Innovation
- Improved efficiency & effectiveness
- New knowledge from combined data sources and patterns in large data volumes

Digital **Scholarship** Services



The Mayor's Office for Economic Opportunity **NYC Opportunity**

Italiano | Translate | Text-Size

About | Poverty in NYC | What We Do | Portfolio | Reports | News | Search

Poverty Measure | Poverty Tool

Share

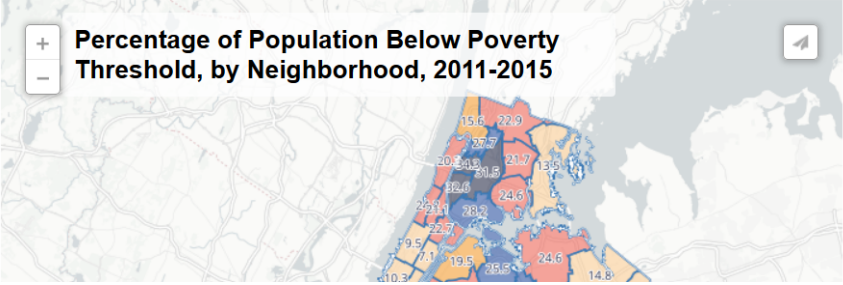
Print

Poverty in NYC

NYC Opportunity is committed to the use of data and evidence in formulating poverty reduction policies. Within NYC Opportunity, the Poverty Research Team generates the alternative poverty measure for New York City, a keystone in innovative, rigorous data analysis.

The Poverty Research Team applies data analytics to build an accurate description of who is in poverty, identify some of the leading causes for being in poverty, and measuring how citywide programs work to offset the poverty rate. This data allows us to better target anti-poverty initiatives and design more effect metrics in measuring success.

Use the map to explore the percentage of the population below the poverty threshold in NYC.



Reproducibility

The Three R's

Image Source: http://merchinsider.com/dealing-with-copycats/

Reproducible
A measurement is reproducible if the investigation is repeated by another person, or by using different equipment or techniques, and the same results are obtained. N.B. "the same" results implies identical, but in reality "the same" means that random error will still be present in the results.

Replication
The ability to independently achieve non identical conclusions that are at least similar, when differences in sampling, research procedures and data analysis methods may exist.

Repeatability
Or test–retest reliability is the variation in measurements taken by a single person or instrument on the same item, under the same conditions, and in a short period of time. A less-than-perfect test–retest reliability causes test–retest variability.

Over half of psychology studies fail reproducibility test

Largest replication study to date casts doubt on many published positive results.

Monya Baker

27 August 2015

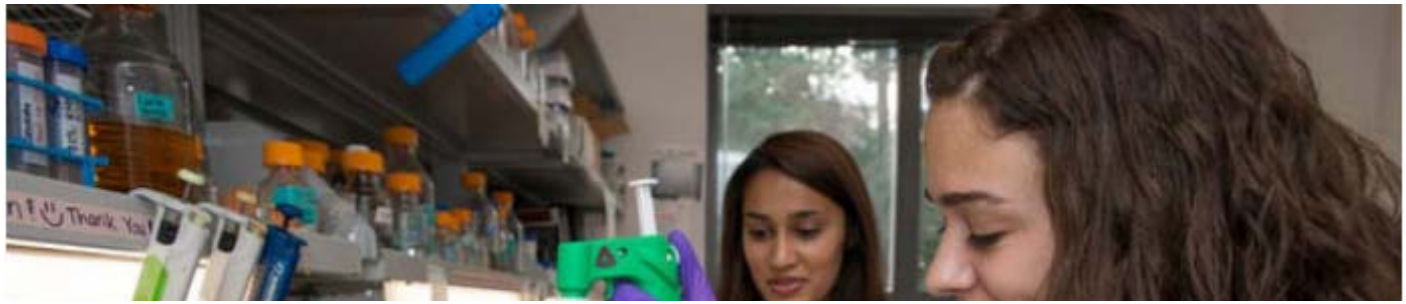
According to the replicators' qualitative assessments, as previously reported by *Nature*, only 39 of the 100 replication attempts were successful. (There were 100 completed replication attempts on the 98 papers, as in two cases replication efforts were duplicated by separate teams.) But literature. In fact, two thirds of it should probably be distrusted.

In the biggest project of its kind, Brian Nosek, a social psychologist and head of the Center for Open Science in Charlottesville, Virginia, and 269 co-authors repeated work reported in 98 original papers from three psychology journals, to see if

Goal of Scientific Publishing

- Two goals:
  - To announce a result and
  - To convince readers that the result is correct

Researchers also receive intellectual credit, recognition, and prestige



You can download our code from the URL supplied. Good luck downloading the only postdoc who can get it to run, though #overlyhonestmethods

-- Ian Holmes (@ianholmes) January 8, 2013

Why it sometimes goes wrong

- Pressure to publish
- Focus on impact factor
- Tainted resources
- Bad math
- Omission
- Messy science
- Issues with peer review
- Some researchers don’t share
- Some research is never shared
- Poor training -> sloppiness
- Honest error
- Fraud
- Disorganization/time pressure
- Cost & time to prepare and curate materials
- Unreplicable data (one-off data, specialist equipment, stochastic

- Transparency
- Replicability
- Triangulation

Correspondence

Preclinical data: Three-point plan for reproducibility

E. Andrew Balas & Lee M. Ellis

The success of clinical studies depends on the reproducibility of preclinical research results (see J. Kimmelman and C. Federico *Nature* **542**, 25–27; 2017). We propose a three-tier principle of transparency, replication and triangulation that should be achieved before publication, to ensure that the results warrant further study in preclinical and clinical trials.

Info | Sections

Subjects: Research data, Research management

Journal: *Nature* **543**, 40 (02 March 2017)

DOI: doi:10.1038/543040d

Download Citation

Published online: 01 March 2017

Transparency
Focuses on the availability of complete and clear information about experimental methodology. This must be sufficient to allow the published study to be replicated under the same conditions by other investigators, with essentially the same primary outcomes.

Replicability
Should be tested by the original researchers and/or by others in the same laboratory, and confirmed using different samples or specimens. Ideally, an unrelated lab should perform independent replication based on the reported methods.

Triangulation
Confirms the study's central findings or models using different methodologies and experiments, ensuring that measurements converge from different experimental perspectives.

What would someone need in order to find, understand, evaluate, and reuse your data?

What is RDM?

- Planning and description
- Document
- Store
- Deposit
- Link

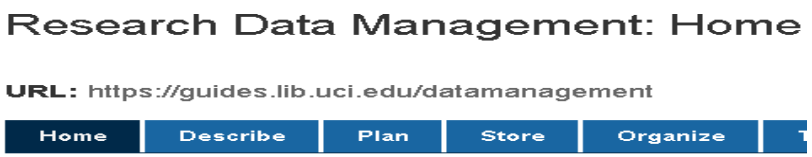
It takes a village…

- Hybrid activity:
  - Researchers
  - Research support personnel
  - Other institutions, commercial partners, etc.

Why you should consider RDM

- Accessibility
- Transparency and quality
- Efficiency
- Speed
- Impact

Digital Scholarship Services (DSS) Resources



https://guides.lib.uci.edu/datamanagement

dash | UCI

dash.lib.uci.edu

UNIVERSITY OF CALIFORNIA EZID *Identifiers made easy*

https://ezid.cdlib.org/

DMPTool

https://dmptool.org/

How DSS can help!

Provide assistance with:

- Writing grant winning Data Management Plans
- Depositing data into repositories for access and preservation
- Capturing metadata to allow re-use
- Creating permanently resolvable hyperlinks
- Connecting your data with your publications

More Information

- Transparency and Openness Promotion (TOP) Guidelines -https://cos.io/our-services/top-guidelines/
- Replicability vs. reproducibility — or is it the other way around? - http://languagelog.ldc.upenn.edu/nll/?p=21956
- F.A.I.R - https://www.force11.org/group/fairgroup/fairprinciples
- HIPAA Security Rule: https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html
- Complying with FERPA: http://dataqualitycampaign.org/wp-content/uploads/2016/03/Complying-with-FERPA-03.2013.pdf
- Digital Scholarship Services – https://www.lib.uci.edu/dss
- Research Data Management Guide - https://guides.lib.uci.edu/datamanagement

Need DSS?

https://www.lib.uci.edu/dss

libdss@uci.edu

Follow us on Twitter #UCILIBDSS

Schedule DSS Consultation