Gaining an Advantage by Sharing Your Research Data

Andrew Sallans
Head of Strategic Data Initiatives
University of Virginia Library
als9q@virginia.edu

Bill Corey
Data Management Consultant
University of Virginia Library
wtc2h@virginia.edu

Goals for the workshop

- Learn about gaining an advantage
- Learn about available resources
- Identify the best places to share your data
- Develop a citation for your dataset
- Gain peer and expert feedback

Why should you care about data sharing?

- It's good science: reproducible results and continuity
- Increases transparency and quality of science
- Published data can be indexed and made discoverable
- Get credit by making your data citable, more impact
- Duplication of data-collecting efforts are reduced
- Increased potential for interdisciplinary research
- You may be required to by your government, funder, institution, publishers, etc.

Dissemination & Sharing of Research Results

"Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing."

National Science Foundation: *Award & Administration Guide (AAG) Chapter VI.D.4*

Recent news

- White House, Office of Science and Technology Policy from February 22, 2013
- Federal research agencies funding more than $100M/year of research and development must develop plan to make the results (papers and data) of federally funded research available to the public within one year of publication
- Promotes the preservation of data and deposit of data in publicly accessible databases and repositories

http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Caveat: it’s not just the NSF

CDC	NEH
DOE	NIH
EPA	USDA
IMLS	Private and public foundations
NASA	Many research funding agencies in the U.K., Australia, and other countries
NOAA	Etc…

Read calls for proposals carefully and ask program director about specific data sharing requirements. Build time into your proposal development to formulate a data management plan that includes data sharing!

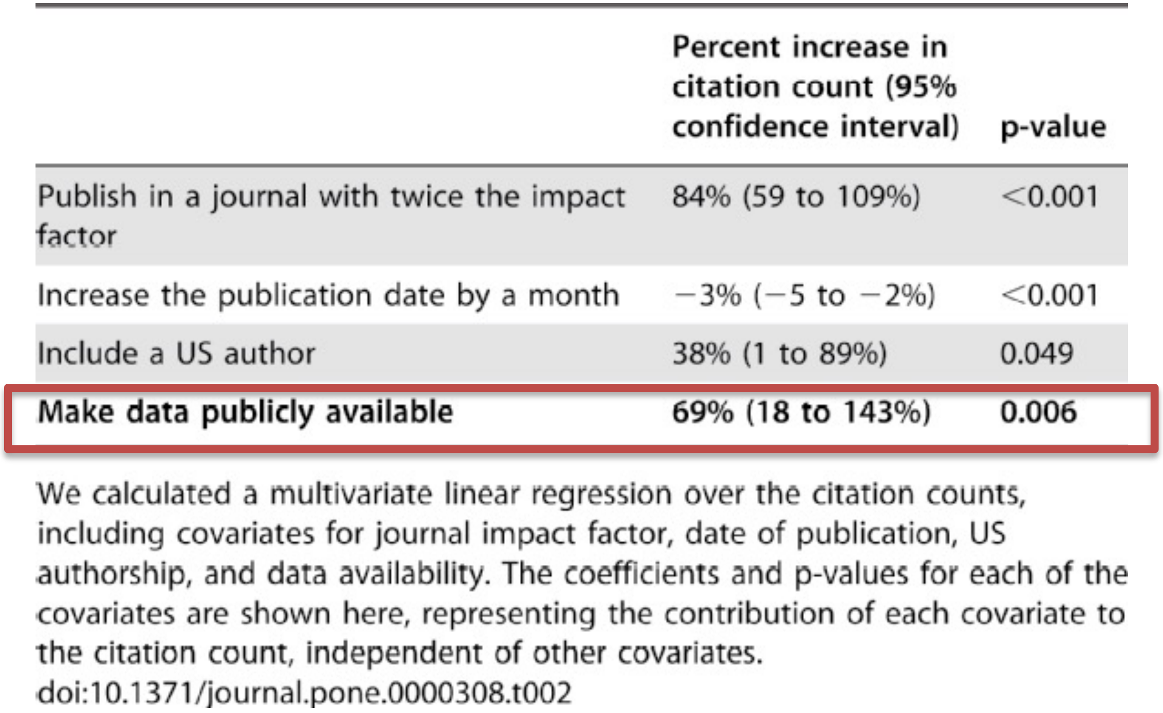
Why not?

UNIVERSITY of VIRGINIA LIBRARY

http://memegenerator.net/Fist-Pump-Baby

Piwowar study, 2007
impact on clinical trial citations

- Examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data
- 48% of trials with publicly available microarray data received 85% of the aggregate citations
- Publicly available data was significantly (p = 0.006) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin using linear regression.



	Percent increase in citation count (95% confidence interval)	p-value
Publish in a journal with twice the impact factor	84% (59 to 109%)	<0.001
Increase the publication date by a month	−3% (−5 to −2%)	<0.001
Include a US author	38% (1 to 89%)	0.049
Make data publicly available	**69% (18 to 143%)**	**0.006**

We calculated a multivariate linear regression over the citation counts, including covariates for journal impact factor, date of publication, US authorship, and data availability. The coefficients and p-values for each of the covariates are shown here, representing the contribution of each covariate to the citation count, independent of other covariates.
doi:10.1371/journal.pone.0000308.t002

Source: http://www.plosone.org/article/info:doi%2F10.1371%2Fjournal.pone.0000308

Advantage: Altmetrics

- Value of things beyond H-index, grant dollars, standard forms of scholarly measurement
- Examples of value/impact
  - Tenure review successes
  - Effect on more citation of articles
  - Effect on more success in grants or collaboration

Advantage: Brian Nosek's COS/OSF

- Reproducibility and replication
- Need some studies/evidence of studies that have been verified to be wrong or new discoveries via OSF
- Strengthen the quality and credibility of your own research by giving others the opportunity to verify

Advantage: Open access/open data

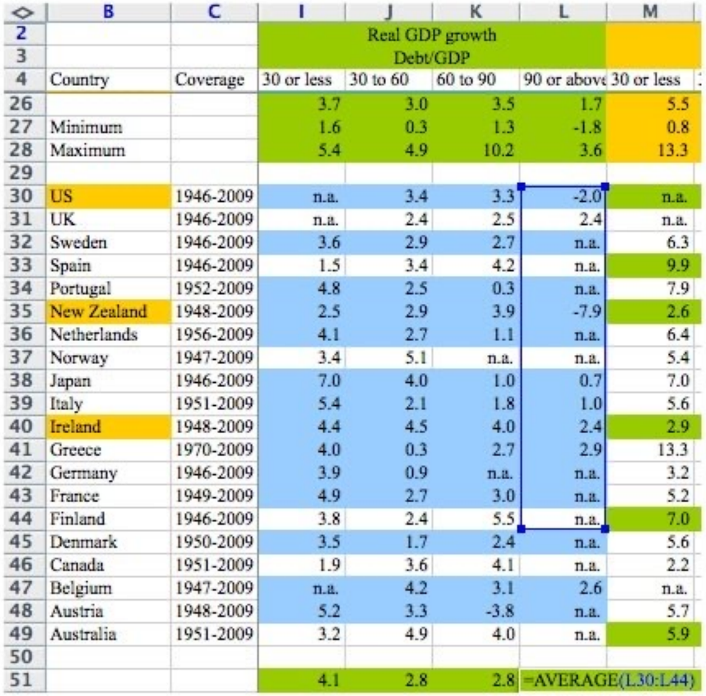
http://nymag.com/daily/intelligencer/2013/04/grad-student-who-shook-global-austerity-movement.html

Advantage: Open access/open data

GROWTH IN A TIME OF DEBT
Carmen M. Reinhart
Kenneth S. Rogoff

Reinhart and Rogoff selectively exclude years of high debt and average growth. Second, they use a debatable method to weight the countries. Third, there also appears to be a coding error that excludes high-debt and average-growth countries.

This is also good evidence for why you should release your data online, so it can be probably vetted.



◇	B	C	I	J	K	L	M
2			Real GDP growth				
3			Debt/GDP				
4	Country	Coverage	30 or less	30 to 60	60 to 90	90 or abov	30 or less
26			3.7	3.0	3.5	1.7	5.5
27	Minimum		1.6	0.3	1.3	-1.8	0.8
28	Maximum		5.4	4.9	10.2	3.6	13.3
29							
30	US	1946-2009	n.a.	3.4	3.3	-2.0	n.a.
31	UK	1946-2009	n.a.	2.4	2.5	2.4	n.a.
32	Sweden	1946-2009	3.6	2.9	2.7	n.a.	6.3
33	Spain	1946-2009	1.5	3.4	4.2	n.a.	9.9
34	Portugal	1952-2009	4.8	2.5	0.3	n.a.	7.9
35	New Zealand	1948-2009	2.5	2.9	3.9	-7.9	2.6
36	Netherlands	1956-2009	4.1	2.7	1.1	n.a.	6.4
37	Norway	1947-2009	3.4	5.1	n.a.	n.a.	5.4
38	Japan	1946-2009	7.0	4.0	1.0	0.7	7.0
39	Italy	1951-2009	5.4	2.1	1.8	1.0	5.6
40	Ireland	1948-2009	4.4	4.5	4.0	2.4	2.9
41	Greece	1970-2009	4.0	0.3	2.7	2.9	13.3
42	Germany	1946-2009	3.9	0.9	n.a.	n.a.	3.2
43	France	1949-2009	4.9	2.7	3.0	n.a.	5.2
44	Finland	1946-2009	3.8	2.4	5.5	n.a.	7.0
45	Denmark	1950-2009	3.5	1.7	2.4	n.a.	5.6
46	Canada	1951-2009	1.9	3.6	4.1	n.a.	2.2
47	Belgium	1947-2009	n.a.	4.2	3.1	2.6	n.a.
48	Austria	1948-2009	5.2	3.3	-3.8	n.a.	5.7
49	Australia	1951-2009	3.2	4.9	4.0	n.a.	5.9
50							
51			4.1	2.8	2.8	=AVERAGE(L30:L44)	

http://www.nber.org/papers/w15639.pdf
http://www.businessinsider.com/thomas-herndon-michael-ash-and-robert-pollin-on-reinhart-and-rogoff-2013-4

Advantage: Data paper example

- Get more from the research that you've done but don't have time to publish
- Get credit and claim precedence (include picture of dog marking territory)
- http://www.ncbi.nlm.nih.gov/pubmed/22373175

Data Papers continued

- Get credit from what somebody else discovers in your data
- “The coolest thing to do with your data will be thought of by someone else.”

  Rufus Pollock

  Cambridge University and Open Knowledge Foundation

- “Open Data is a philosophy and practice requiring that certain data are freely available to everyone, without restrictions from copyright, patents or other mechanisms of control.”

  Peter Murray-Rust

  Cambridge University (in Wikipedia)

Advantage: Data Journal

A data journal is an open access platform where data can be published, peer-reviewed, and cited.

- Evidence of data journals impact

MUST FOLLOW POLICY

- UVA policy
- Funder policy
- See our guidance document
- JoRD
- Include stop sign image

Resources: Figshare/repositories

Sharable, discoverable, and citable

- Beautify this stuff, show logos, show stats of use, show downsteam impact
- http://www.plumanalytics.com/pr/plum-analytics-and-figshare-collaborate.html
- DataCite saw highest stats for Figshare items: http://www.datacite.org/node/76

Resources: Impact Story

- Beautify this stuff, show logos, show stats of use, show downsteam impact
- Moving from “raw altmetrics data to data-driven stories” http://blog.impactstory.org/2012/09/14/31524247207/

Resources: licensing basics you must know

- Data can not be copyrighted.
- Database structure and data sheets can be copyrighted.
- Database content may be copyrighted if viewed as a collection or compilation.
- Intellectual Property rights in data may prevent third-parties from using your data without explicit permission.

Resources: licensing how-to

A license reduces uncertainty about re-use.

- Publically-funded research data should use a PDDL (Public Domain Dedication License) or CC0 (Creative Commons Zero) license to avoid ‘attribution stacking’.
- Creative Commons: creativecommons.org/licenses/
- Open Data Commons: opendatacommons.org/licenses/
- GNU software licenses http://www.gnu.org/licenses/

Resources: DOIs why/how-to

- Benefits of DOIs
- Stats on DOIs
- How to DOI

Resources: DOIs how-to

- More than 98% of all DOI registered are for scholarly articles
- Creator (PublicationYear): Title. Publisher. Identifier
- Example: 10.1594/PANGAEA.484677
- Citations for this DOI   doi:10.1594/PANGAEA.484677
- hypertext link: http://dx.doi.org/10.1594/PANGAEA.484677
- Resolves to:
  http://doi.pangaea.de/10.1594/PANGAEA.484677

http://www.datacite.org/whatisdoi

Resources: DOIs why

- A DOI Name (DOI) is a specific type of Handle and can be assigned to any object that is a form of intellectual property
- The assignment of a DOI Name indicates that a dataset will be well managed and accessible for long-term use.
- Persistent citations in scholarly materials (journal articles, books, etc.) through CrossRef, a consortium of around 3,000 publishers
- Scientific data sets, through DataCite, a consortium of leading research libraries, technical information providers, and scientific data centers

Team Exercise
30 minutes

1. Identify the options for where to share your dataset.
2. Evaluate the benefits of each.
3. Review requirements for sharing location and prepare dataset to share.
4. Specify how your data should be cited.
5. Record issues and questions for discussion.

Presentation of Dataset Share
15 minutes

- Identify options evaluated
- Describe decisions briefly
- Explain requirements
- Describe advantages to final choice
- Show final citation

Questions and Discussion?

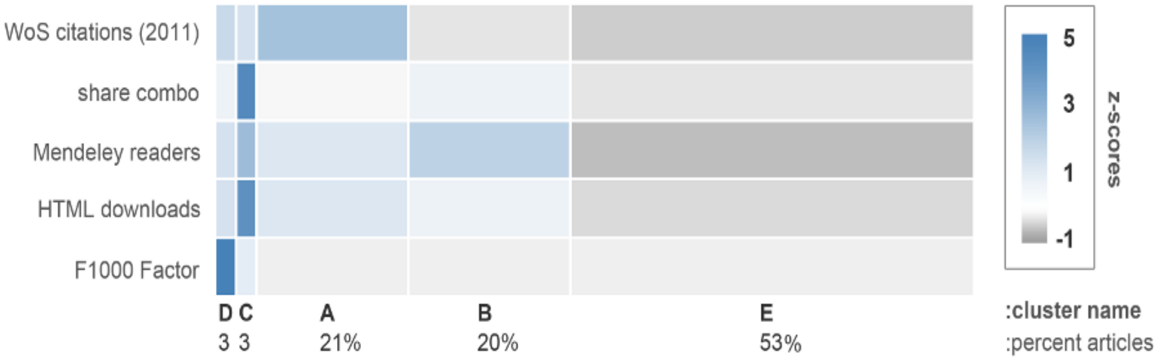
Follow-up

- Contact the Data Management Consulting Group for help with data sharing
- Email: DMConsult@virginia.edu

Advantage: Altmetrics

Value of things beyond H-index, grant dollars, standard forms of scholarly measurement

We find that that different indicators vary greatly in activity. Around 5% of sampled articles are cited in Wikipedia, while close to 80% have been included in at least one Mendeley library. There is, however, an encouraging diversity; a quarter of articles have nonzero data from five or more different sources. Correlation and factor analysis suggest citation and altmetrics indicators track related but distinct impacts with neither able to describe the complete picture of scholarly use alone. There are moderate correlations between Mendeley and Web of Science citation, but many altmetric indicators seem to measure impact mostly orthogonal to citation. Articles cluster in ways that suggest five different impact "flavors", capturing impacts of different types on different audiences; for instance, some articles may be heavily read and saved by scholars but seldom cited.



http://arxiv.org/abs/1203.4745v1

Strategy to figure out whether this stuff actually matters

Questions for Discussion

- Statistics on impact and benefit to adopters
- Have they seen examples of people incorporating into grant broader impacts, CVs, etc.?
- Have they seen examples of people using in tenure and promotion review?