Prepared for

US DATA CITATION WORKSHOP:
DEVELOPING POLICY AND PRACTICE

July 2016

The National Academies of Sciences, Engineering, and Medicine

Rewards and Incentives from a Researcher's Perspective

Dr. Micah Altman
<escience@mit.edu>

Director of Research, MIT Libraries

DISCLAIMER

These opinions are my own, they are not the opinions of MIT, any of the project funders, nor (with the exception of co-authored previously published work) my collaborators

Secondary disclaimer:

"It's tough to make predictions, especially about the future!"

-- Attributed to Woody Allen, Yogi Berra, Niels Bohr, Vint Cerf, Winston Churchill, Confucius, Disreali [sic], Freeman Dyson, Cecil B. Demille, Albert Einstein, Enrico Fermi, Edgar R. Fiedler, Bob Fourer, Sam Goldwyn, Allan Lamport, Groucho Marx, Dan Quayle, George Bernard Shaw, Casey Stengel, Will Rogers, M. Taub, Mark Twain, Kerr L. White, etc.

Related Publications

- Altman, M., Wood, A., O'Brien, D.R., Vadhan, S. and Gasser, U., 2015. Towards a Modern Approach to Privacy-Aware Government Data Releases. *Berkeley Tech. LJ,* 30, pp.1967-2073.
- Smith, Yoshimura, Karen, M. Altman, et al, Registering Researchers in Authority Files, *OCLC* 2014.
- Allen, Liz, Amy Brand, Jo Scott, Micah Altman, and Marjorie Hlava. "Credit where credit is due." *Nature* 508 (2014): 312-313.
- Brand, Amy, Liz Allen, Micah Altman, Marjorie Hlava, and Jo Scott. "Beyond authorship: attribution, contribution, collaboration, and credit." *Learned Publishing* 28, no. 2 (2015): 151-155.
- Altman, Micah, and Mercè Crosas. "The Evolution of Data Citation: From Principles to Implementation." *IASSIST Quarterly* (2013): 63.
- Altman M, Jackman S. "Nineteen Ways of Looking at Statistical Software". Journal of Statistical Software. 2011;42.
- Altman, Micah, and Gary King. "A proposed standard for the scholarly citation of quantitative data." *D-lib* 13, no. 3 (2007): 5.
- Altman, M., Gill, J. and McDonald, M.P., 2004. Numerical issues in statistical computing for the social scientist. John Wiley & Sons.

Reprints available from:
informatics.mit.edu

Today's Perspectives & Provocations

* Principles *
* Progress *
* Perforations *
* Prospects *

Principles

What Motivates Scientists?

- Puzzle-solving
- Recognition
- Money

See: Stephan, Paula E. *How economics shapes science*. Cambridge, MA: Harvard University Press, 2012.

What are Reward Channels?

- Puzzle-solving
  - Collaboration
  - Students and postdocs
  - Lab environment
- Recognition
  - Publication
    - Priority of publication and discovery
    - Reputation of publication outlet
  - Readership and use
  - Citation
  - Professional associations and publication roles
  - Prizes
- Money
  - Hiring and job market
  - Tenure & promotion
  - Grants and awards
  - Consulting, startups, intellectual property

Norms, External Incentives, Compliance

- Policy Leaders
  - Funder
  - Associations
  - University
  - Publisher
- Compliance
  - Compliance relies on tie with external incentive
  - Compliance incents satisficing, not optimization
  - Compliance requires monitoring
  - Compliance may displace norms
- Norms and nudges
  - Engaging internal incentives can be very efficient mechanism
  - Norms vary across communities of practice
  - Can be destroyed by compliance/external incentives
  - Often transmitted through tacit knowledge
  - Difficult to re-establish

See:

Gneezy, U. and Rustichini, A., 2000. Pay enough or don't pay at all. Quarterly journal of economics, pp.791-810.
Ostrom, E., 2009. Understanding institutional diversity. Princeton university press.
Borgman, C.L., 2010. *Scholarship in the digital age: Information, infrastructure, and the Internet*. MIT press..

Inequalities

- Vast differences in productivity across scientists (see for example, Lotka's law)
- Impossible to determine precisely what portions stem from "ability" vs. early resource/attention advantages – however a substantial portion is likely "state dependent"
- Broad and substantial gender disparities exist in science and scientific outputs
- Different supports for mentoring (tacit knowledge) and collaboration; and tenure and promotion practice probably necessary to address disparities
- Contributorship roles and reproducibility practices are important part of mentoring and collaboration activities

See:

Stephan, Paula E. *How economics shapes science*. Cambridge, MA: Harvard University Press, 2012

Sugimoto, C.R., Lariviere, V., Ni, C.Q., Gingras, Y. and Cronin, B., 2013. Global gender disparities in science. Nature, 504(7479), pp.211-213.

Ceci, S.J. and Williams, W.M., 2011. Understanding current causes of women's underrepresentation in science. Proceedings of the National Academy of Sciences, 108(8), pp.3157-3162.

Bozeman, B. and Gaughan, M., 2011. How do men and women differ in research collaborations? An analysis of the collaborative motives and strategies of academic researchers. Research Policy, 40(10), pp.1393-1402.

Progress

The Bad Old Days -- 2007

"**How much slower would scientific progress be if the near universal standards for scholarly citation of articles and books had never been developed?** Suppose shortly after publication only some printed works could be reliably found by other scholars; or if researchers were only permitted to read an article if they first committed not to criticize it, or were required to coauthor with the original author any work that built on the original. How many discoveries would never have been made if the titles of books and articles in libraries changed unpredictably, with no link back to the old title; if printed works existed in different libraries under different titles; if researchers routinely redistributed modified versions of other authors' works without changing the title or author listed; or if publishing new editions of books meant that earlier editions were destroyed? …

"**Unfortunately, no such universal standards exist for citing quantitative data, and so all the problems listed above exist now.** Practices vary from field to field, archive to archive, and often from article to article.

The data cited may no longer exist, may not be available publicly, or may have never been held by anyone but the investigator. Data listed as available from the author are unlikely to be available for long and will not be available after the author retires or dies. Sometimes URLs are given, but they often do not persist. In recent years, a major archive renumbered all its acquisitions, rendering all citations to data it held invalid; identical data was distributed in different archives with different identifiers; data sets have been expanded or corrected and the old data, on which prior literature is based, was destroyed or renumbered and so is inaccessible; and modified versions of data are routinely distributed under the same name, without any standard for versioning. Copyeditors have no fixed rules, and often no rules whatsoever. Data are sometimes listed in the bibliography, sometimes in the text, sometimes not at all, and rarely with enough information to guarantee future access to the identical data set. Replicating published tables and figures even without having to rerun the original experiment, is often difficult or impossible"

See:
Altman, Micah, and Gary King. "A proposed standard for the scholarly citation of quantitative data." D-lib 13, no. 3 (2007):

Now

Office of Science and Technology Policy

Ensure **that all extramural researchers receiving Federal grants and contracts for scientific research** and intramural researchers **develop data management plans**, as appropriate, describing how they will provide for long-term preservation of, and access to, scientific data in digital formats resulting from federally funded research, or explaining why long term preservation and access cannot be justified…

PLOS ONE

All data and related metadata underlying the findings reported in a submitted manuscript should be deposited in an appropriate public repository,
…
The Data Availability Statement **must specify that data are deposited publicly and list the name(s) of repositories along with digital object identifiers or accession numbers** for the relevant data sets.

DC1
Data Citation Principles

Data should be considered legitimate, citable products of research.

The New York Times

SundayReview

Why Do So Many Studies Fail to Replicate?

Some Notable Changes

- Compliance
  - Funder: data management plans, open data
  - Publishers: data access/archiving/citation
- Norms & practices
  - Joint data citation principles
  - Recognition of data in funder biosketches
  - Increased recognition of reproducibility gaps
  - Increased recognition of open data/open science
- Technical infrastructure
  - Open data repositories
  - Data citation indices
  - ORCID researcher identifier and registry
- Recognition
  - Data citation indices
  - Virtual branded archives
  - High-profile data publications

DC1 Data Citation Principles

The Dataverse Project

figshare

DataCite

SCIENTIFIC DATA

Perforations

Limitations of Compliance

- Monitoring
  - Universities are not routinely monitoring data sharing compliance
  - Uncertain how funders will monitor
  - Most data management plans are not publicly published/auditable
  - Compliance with journal open data policies is mixed

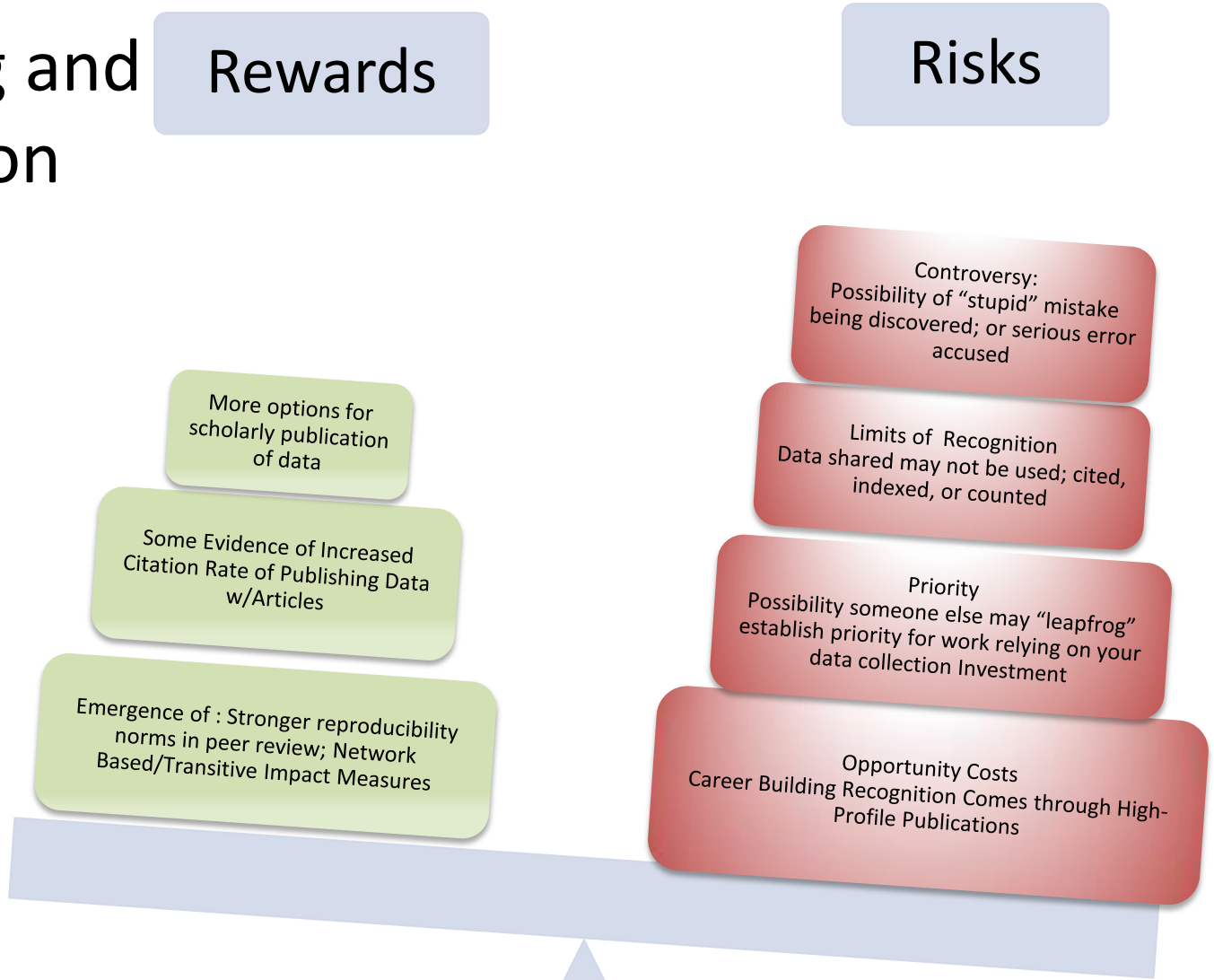
- Uncertain scope

"explaining why long term preservation and access cannot be justified"

  - Lack of standard set of practices for access to big data
  - Tension between open data/reproducibility vs.
    - Commercial data, intellectual property
    - Personal data

See: Altman, M., Wood, A., O'Brien, D.R., Vadhan, S. and Gasser, U., 2015. Towards a Modern Approach to Privacy-Aware Government Data Releases. Berkeley Tech. LJ, 30, pp.1967-2073.
Dimitrova, V., Open Research Data in Economics. Issues in Open Research Data, p.141.2014.
"Enabling Reproducibility in Big Data Research: Balancing Confidentiality and Scientific Transparency," chapter in Lane, J., Stodden, V., Bender, S., and Nissenbaum, H. (eds). 2014. Privacy, Big Data, and the Public Good: Frameworks for Engagement. Cambridge University Press.

Data Sharing and Recognition



See: Katz, D., 2014. Transitive credit as a means to address social and technological concerns stemming from citation and attribution of digital products. *Journal of Open Research Software*, 2(1).
Borgman, C.L., 2015. *Big data, little data, no data: Scholarship in the networked world*. Mit Press.
Nosek, B.A., Spies, J.R. and Motyl, M., 2012. Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. Perspectives on Psychological Science, 7(6), pp.615-631.

Prospects

Bridging Data and Software

- Most published empirical scientific claims rely on both data and software
- Current state of scientific software curation/citation is analogous to data curation/citation in 2006.
- Recent findings on inflated false-positive rates in fMRI inference underscores issue

See: Buckheit, Jonathan B., and David L. Donoho. "Wavelab and reproducible research" Department of Statistics, Stanford U. 1995; Altman, M., Gill, J. and McDonald, M.P., 2004. *Numerical issues in statistical computing for the social scientist*. John Wiley & Sons; Altman M, Jackman S. "Nineteen Ways of Looking at Statistical Software". *Journal of Statistical Software*. 2011;42. ; Eklund, A., Nichols, T.E. and Knutsson, H., 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*,

Initiatives

- FORCE 11 Software Citation Principles

  www.force11.org/software-citation-principles

- ACM New Publication Policies on Software Reproducibility and Contributorship

  www.acm.org/publications/policies

- Software Preservation:

  - www.softwarepreservationnetwork.org
  - www.softwareheritage.org
  - guides.github.com/activities/citable-code/

Recognizing Contributor Roles

Contributor Roles/Data curation

Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later re-use.

- Most published empirical scientific claims rely on both data and software
- Current state of scientific software curation/citation is analogous to data curation/citation in 2006.

See: Allen, Liz, Amy Brand, Jo Scott, Micah Altman, and Marjorie Hlava. "Credit where credit is due." *Nature* 508 (2014): 312-313.
Brand, Amy, Liz Allen, Micah Altman, Marjorie Hlava, and Jo Scott. "Beyond authorship: attribution, contribution, collaboration, and credit." *Learned Publishing* 28, no. 2 (2015): 151-155.

Initiatives

- Standardization

casrai.org/credit

- Systems Integration
  www.ariessys.com/software/standards/
- Authorship Policies
  - PLOS: journals.plos.org/plosone/s/authorship
  - CELL: www.cell.com/cell/authors
  - ACM: www.acm.org/publications/policies/policy_on_authorship

Questions?

E-mail: escience@mit.edu

Web: informatics.mit.edu