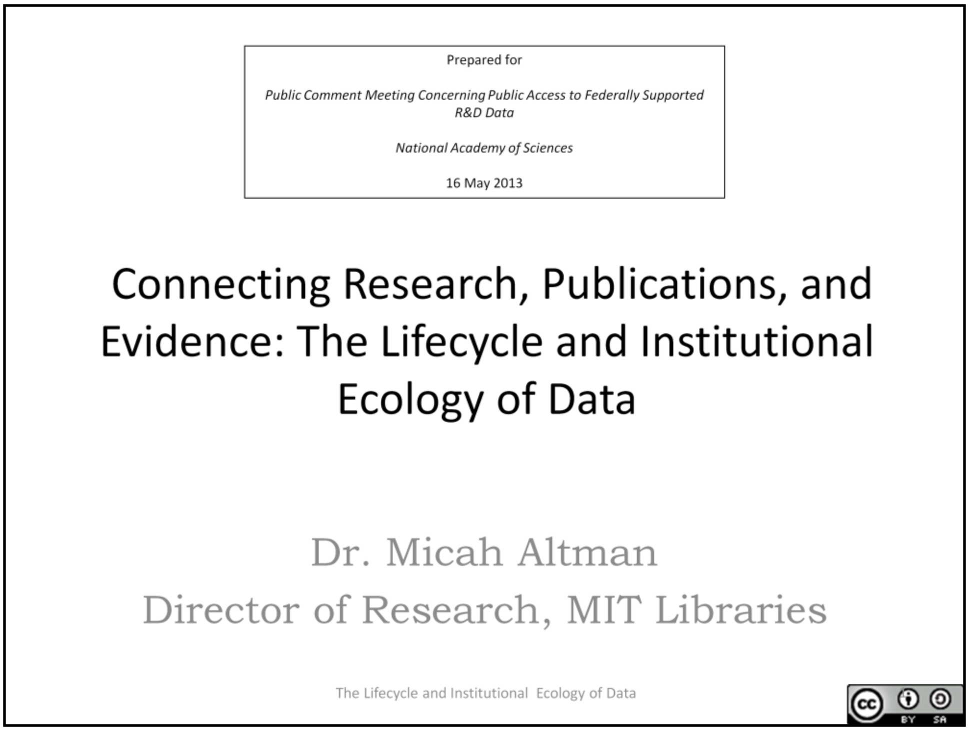


Prepared for

Public Comment Meeting Concerning Public Access to Federally Supported R&D Data

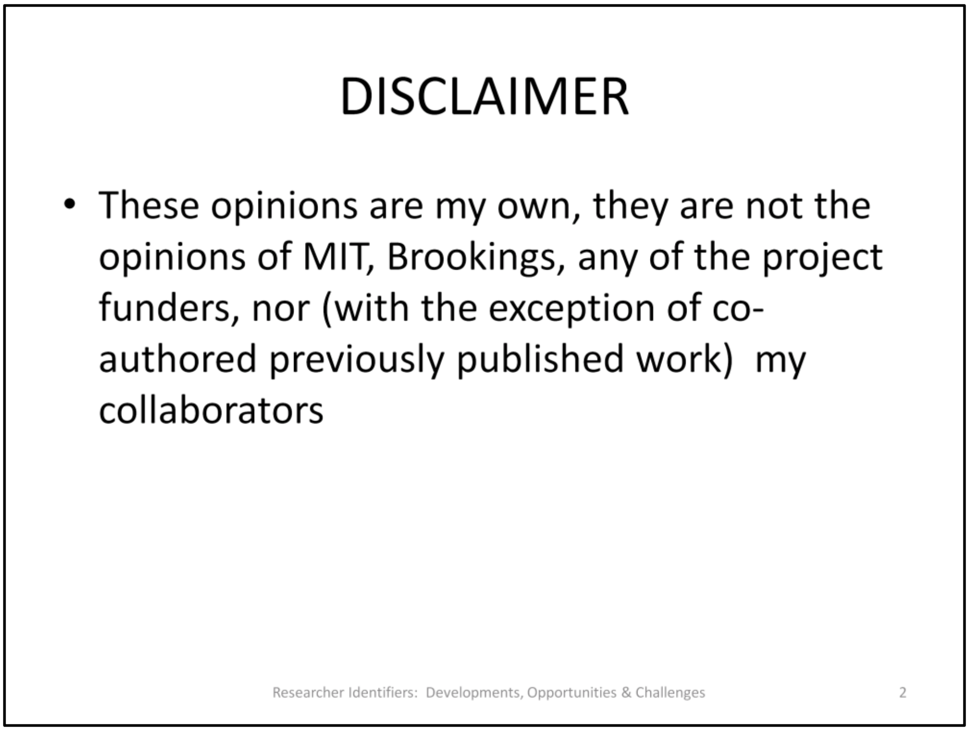
National Academy of Sciences

16 May 2013

Connecting Research, Publications, and Evidence: The Lifecycle and Institutional Ecology of Data

Dr. Micah Altman
Director of Research, MIT Libraries

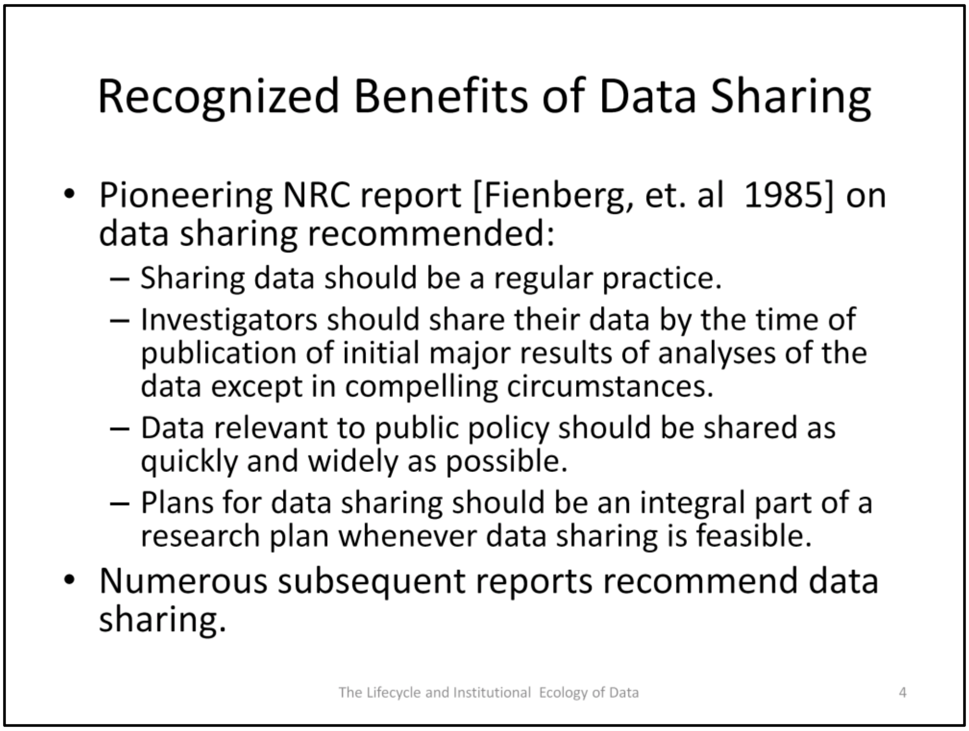
The Lifecycle and Institutional Ecology of Data



DISCLAIMER

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

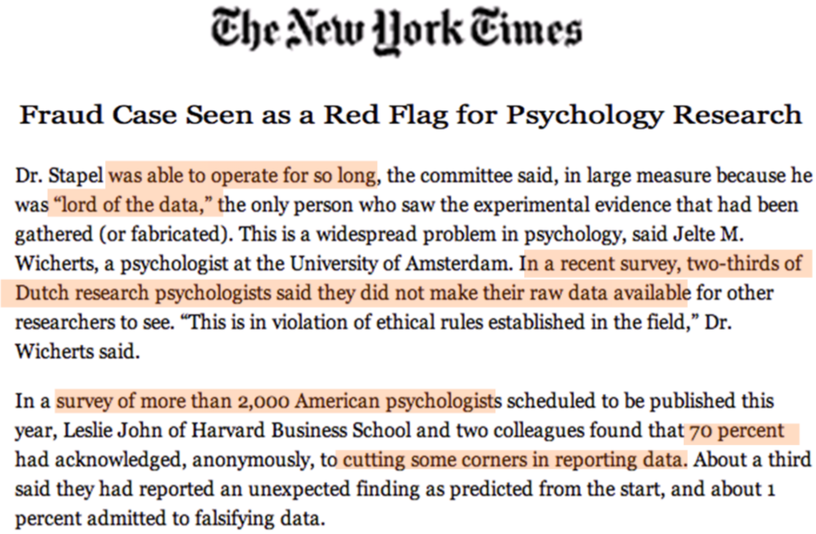
Motivations



Recognized Benefits of Data Sharing

- Pioneering NRC report [Fienberg, et. al 1985] on data sharing recommended:
  - Sharing data should be a regular practice.
  - Investigators should share their data by the time of publication of initial major results of analyses of the data except in compelling circumstances.
  - Data relevant to public policy should be shared as quickly and widely as possible.
  - Plans for data sharing should be an integral part of a research plan whenever data sharing is feasible.
- Numerous subsequent reports recommend data sharing.

Increased Retractions, Allegations of Fraud



The New York Times

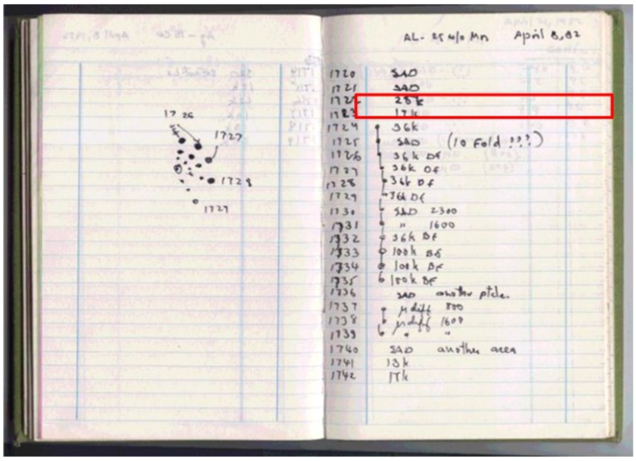
Fraud Case Seen as a Red Flag for Psychology Research

Dr. Stapel was able to operate for so long, the committee said, in large measure because he was "lord of the data," the only person who saw the experimental evidence that had been gathered (or fabricated). This is a widespread problem in psychology, said Jelte M. Wicherts, a psychologist at the University of Amsterdam. In a recent survey, two-thirds of Dutch research psychologists said they did not make their raw data available for other researchers to see. "This is in violation of ethical rules established in the field," Dr. Wicherts said.

In a survey of more than 2,000 American psychologists scheduled to be published this year, Leslie John of Harvard Business School and two colleagues found that 70 percent had acknowledged, anonymously, to cutting some corners in reporting data. About a third said they had reported an unexpected finding as predicted from the start, and about 1 percent admitted to falsifying data.

Unpublished Data Ends up in the "Desk Drawer"

- Null results are less likely to be published →
  published results as a whole are biased toward positive findings
- Outliers are routinely discarded →
  unexpected patterns of evidence across studies remain hidden



AL- 25 w/o Mn April 8, 82

1720	SAD	
1721	SAD	
1722	25k	
1723	17k	
1724	36k	
1725	SAD	(10 Fold ???)
1726	36k DF	
1727	36k DF	
1728	36k DF	
1729	36k DF	
1730	SAD	2300
1731	"	1600
1732	36k DF	
1733	100k BF	
1734	100k BF	
1735	100k BF	
1736	SAD	another ptcle.
1737	μ diff	800
1738	μ diff	1600
1739	"	"
1740	SAD	another area
1741	13k	
1742	17k	

Daniel Schectman's Lab Notebook Providing Initial Evidence of Quasi Crystals

Erosion of Evidence Base

- Researchers lack archiving capability
- Individualized incentives for preserving evidence base are weak

Examples

Intentionally Discarded: “Destroyed, in accord with [nonexistent] APA 5-year post-publication rule.”

Unintentional Hardware Problems “Some data were collected, but the data file was lost in a technical malfunction.”

Acts of Nature The data from the studies were on punched cards that were destroyed in a flood in the department in the early 80s.”

Discarded or Lost in a Move “As I retired …. Unfortunately, I simply didn’t have the room to store these data sets at my house.”

Obsolescence “Speech recordings stored on a LISP Machine…, an experimental computer which is long obsolete.”

Simply Lost “For all I know, they are on a [University] server, but it has been literally years and years since the research was done, and my files are long gone.”

Research by:
ICPSR

4

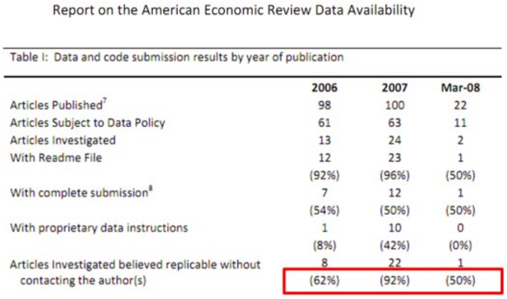
Compliance with Data Sharing Policies is often Low

- Compliance is low even in best examples of journals
- Checking compliance is labor-intensive without citation and repository standards

The American Economic Review

Papers in Honor of the Centenary of the *AER*

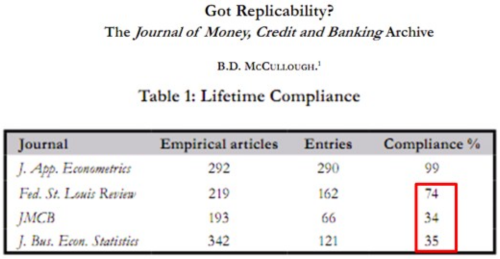
FEBRUARY 2011



Report on the American Economic Review Data Availability

Table I: Data and code submission results by year of publication

	2006	2007	Mar-08
Articles Published[7]	98	100	22
Articles Subject to Data Policy	61	63	11
Articles Investigated	13	24	2
With Readme File	12 (92%)	23 (96%)	1 (50%)
With complete submission[8]	7 (54%)	12 (50%)	1 (50%)
With proprietary data instructions	1 (8%)	10 (42%)	0 (0%)
Articles Investigated believed replicable without contacting the author(s)	8 (62%)	22 (92%)	1 (50%)

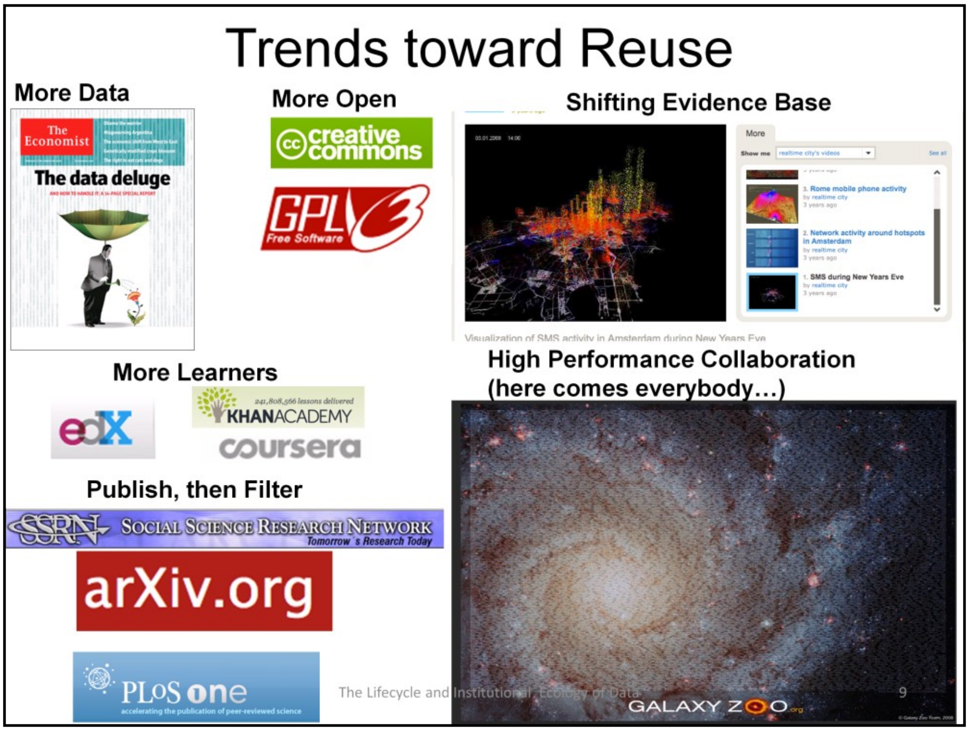


Got Replicability?
The *Journal of Money, Credit and Banking* Archive

B.D. McCullough.[1]

Table 1: Lifetime Compliance

Journal	Empirical articles	Entries	Compliance %
J. App. Econometrics	292	290	99
Fed. St. Louis Review	219	162	74
JMCB	193	66	34
J. Bus. Econ. Statistics	342	121	35



Trends toward Reuse

More Data

The Economist

The data deluge

More Open

creative commons

GPL 3 Free Software

Shifting Evidence Base

3. Rome mobile phone activity
by realtime city
3 years ago

2. Network activity around hotspots in Amsterdam
by realtime city
3 years ago

1. SMS during New Years Eve
by realtime city
3 years ago

Visualization of SMS activity in Amsterdam during New Years Eve

More Learners

edX

241,808,566 lessons delivered KHANACADEMY

coursera

High Performance Collaboration (here comes everybody…)

GALAXY ZOO.org

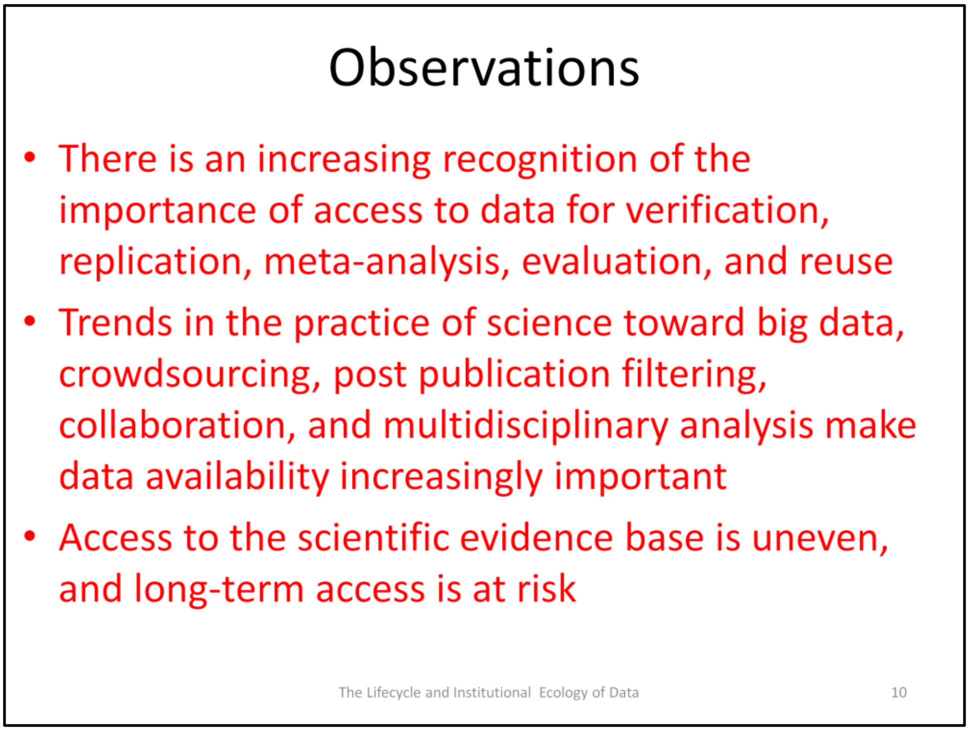
Publish, then Filter

SSRN Social Science Research Network Tomorrow's Research Today

arXiv.org

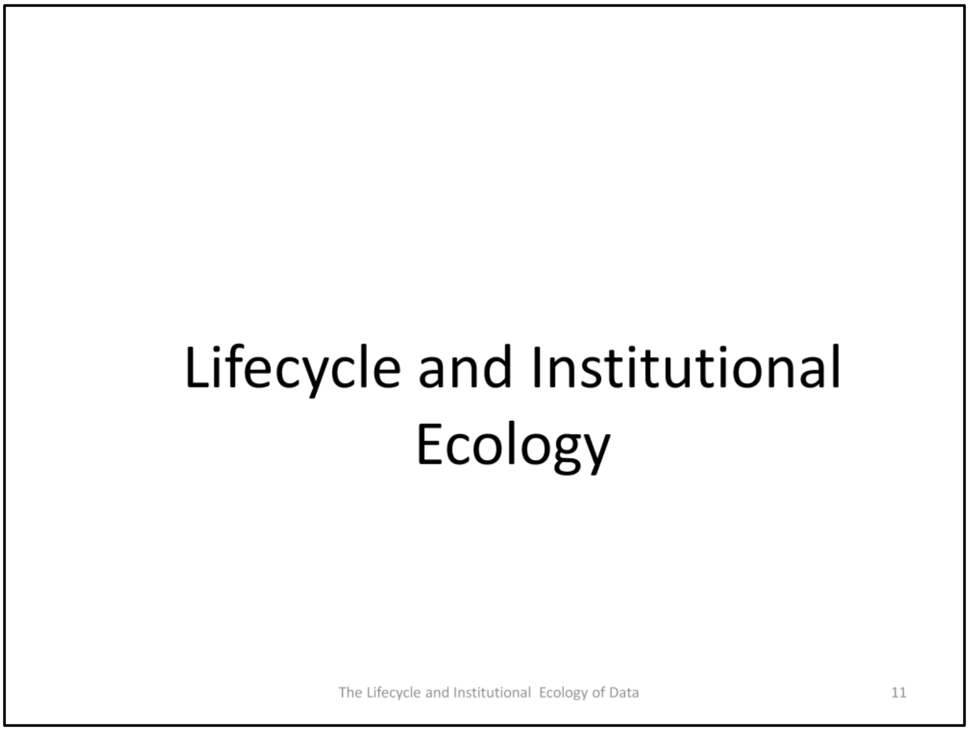
PLoS one accelerating the publication of peer-reviewed science

LHC produces a PB every 2 weeks, Sloan Galaxy zoo has hundreds of thousands of “authors”, 50K people attend a class from the University of michigan, and to understand public opinion instead of surveying 100’s of people per month we can analyze 10ooo tweets per second.

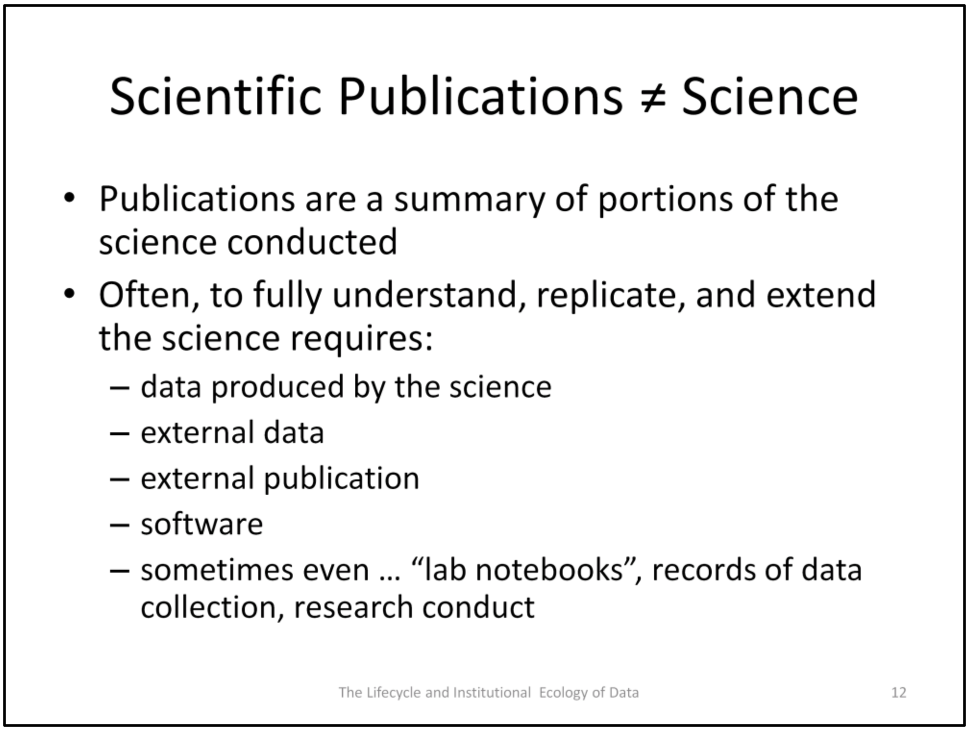


Observations

- There is an increasing recognition of the importance of access to data for verification, replication, meta-analysis, evaluation, and reuse
- Trends in the practice of science toward big data, crowdsourcing, post publication filtering, collaboration, and multidisciplinary analysis make data availability increasingly important
- Access to the scientific evidence base is uneven, and long-term access is at risk

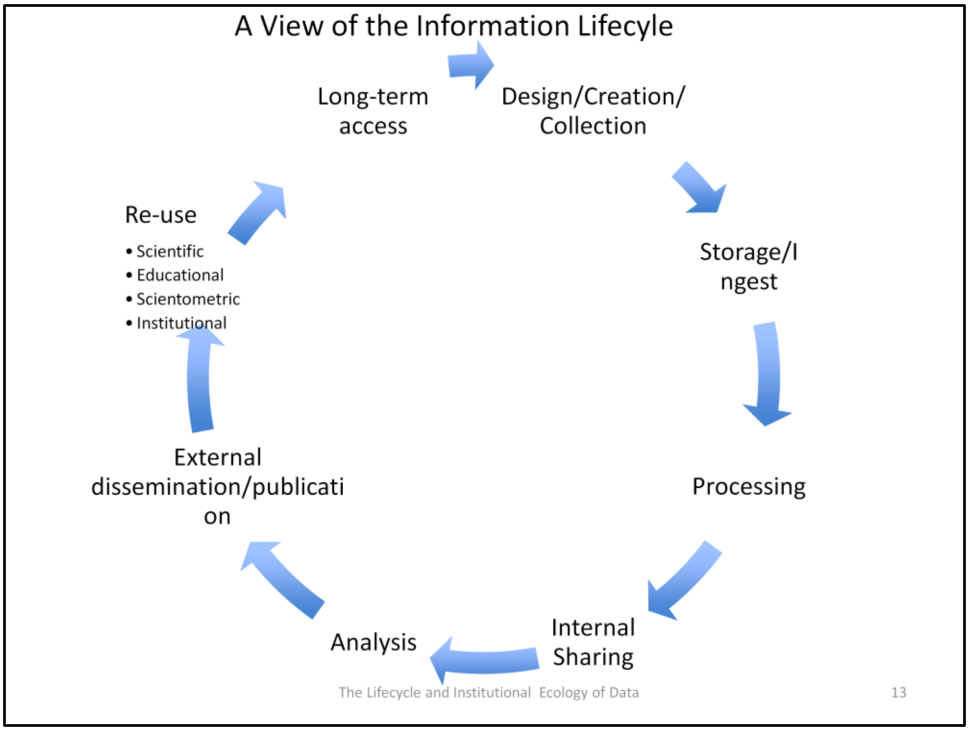


Lifecycle and Institutional Ecology



Scientific Publications ≠ Science

- Publications are a summary of portions of the science conducted
- Often, to fully understand, replicate, and extend the science requires:
  - data produced by the science
  - external data
  - external publication
  - software
  - sometimes even … “lab notebooks”, records of data collection, research conduct



A View of the Information Lifecyle

Design/Creation/ Collection

Storage/Ingest

Processing

Internal Sharing

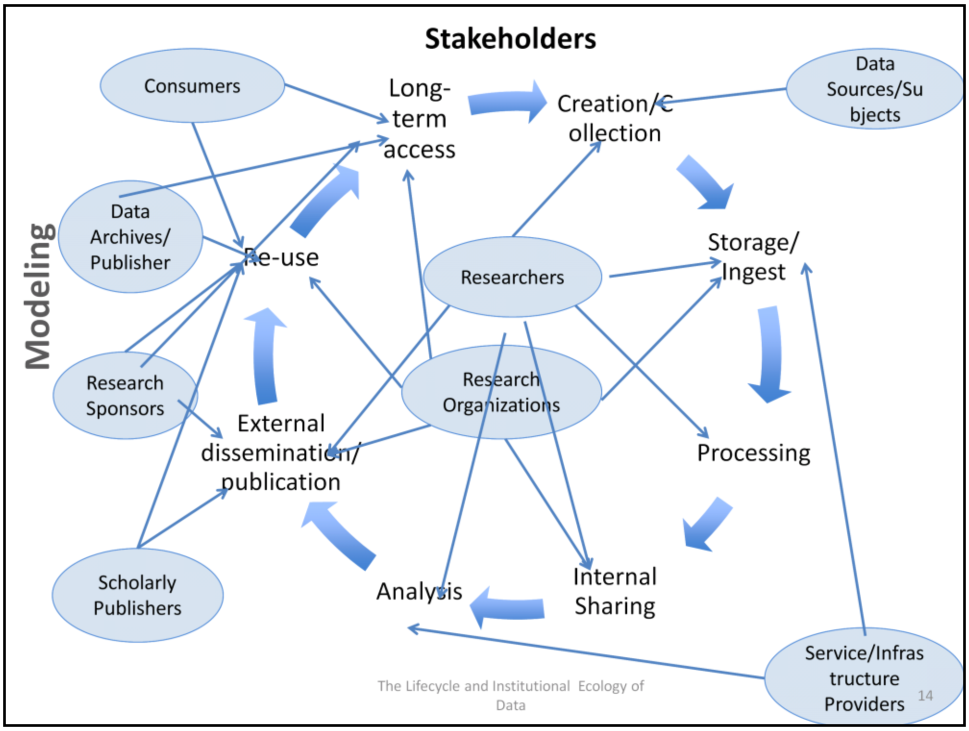
Analysis

External dissemination/publication

Re-use

- Scientific
- Educational
- Scientometric
- Institutional

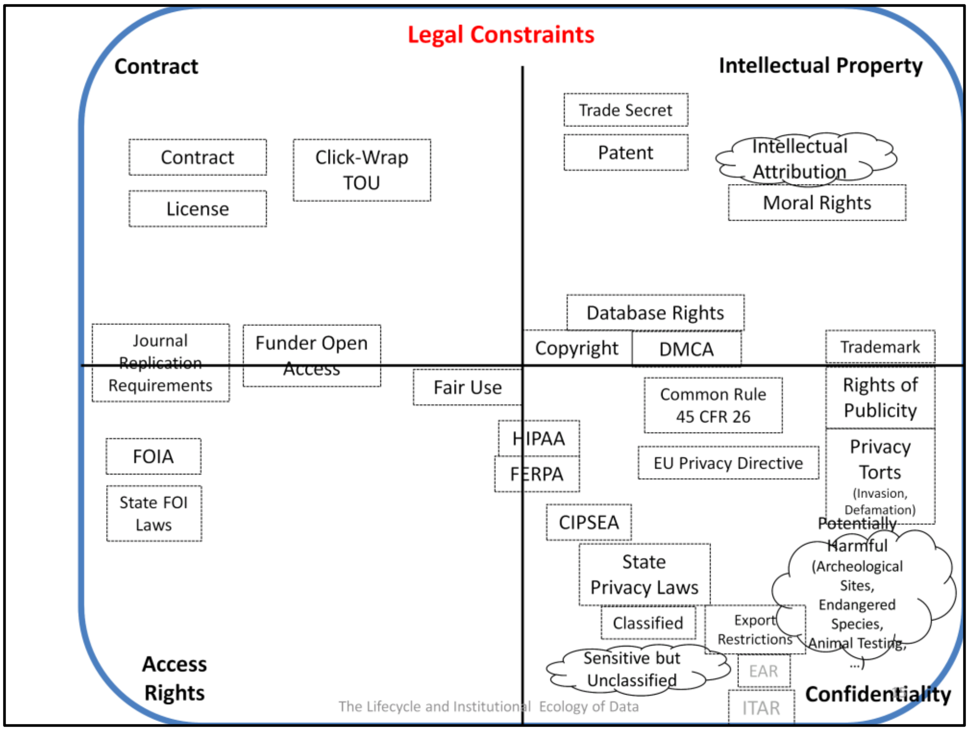
Long-term access



Most of the different stakeholders have stronger relationships/stakes with research at different stages.

But researchers and research institutions are in the middle – they have a strong stake in most stages

Researchers are more directly concerned with collection, processing, analysis, dissemination. Organizations have a higher stake in internal sharing, re-use, long-term access.



Legal Constraints

Contract

Intellectual Property

Contract

Click-Wrap TOU

License

Trade Secret

Patent

Intellectual Attribution

Moral Rights

Database Rights

Journal Replication Requirements

Funder Open Access

Fair Use

Copyright

DMCA

Trademark

Rights of Publicity

Privacy Torts (Invasion, Defamation)

Common Rule 45 CFR 26

EU Privacy Directive

HIPAA

FERPA

FOIA

State FOI Laws

CIPSEA

State Privacy Laws

Potentially Harmful (Archeological Sites, Endangered Species, Animal Testing, …)

Classified

Export Restrictions

EAR

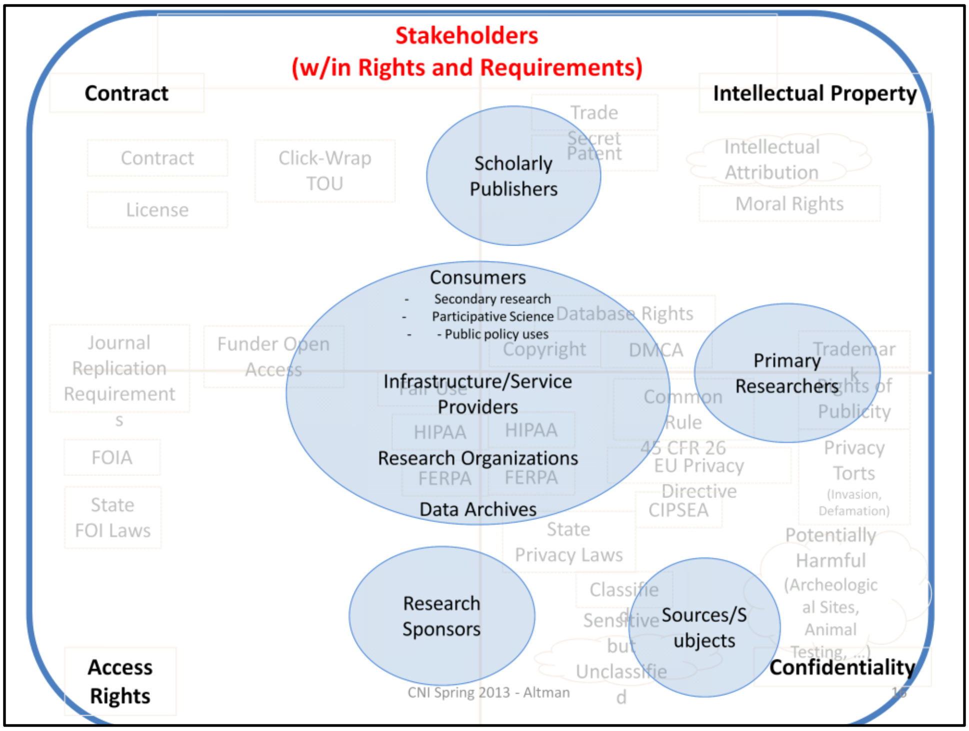
ITAR

Sensitive but Unclassified

Access Rights

Confidentiality

The Lifecycle and Institutional Ecology of Data



Stakeholders
(w/in Rights and Requirements)

Contract

Contract

License

Click-Wrap TOU

Scholarly Publishers

Trade Secret

Patent

Intellectual Property

Intellectual Attribution

Moral Rights

Consumers
- Secondary research
- Participative Science
- - Public policy uses

Database Rights

Copyright

DMCA

Fair Use

Journal Replication Requirements

Funder Open Access

Infrastructure/Service Providers

HIPAA

HIPAA

Research Organizations

FERPA

FERPA

Data Archives

Primary Researchers

Trademark

Rights of Publicity

Common Rule

45 CFR 26

EU Privacy Directive

CIPSEA

Privacy Torts (Invasion, Defamation)

FOIA

State FOI Laws

State Privacy Laws

Potentially Harmful (Archeological Sites, Animal Testing, …)

Research Sponsors

Classified

Sensitive but Unclassified

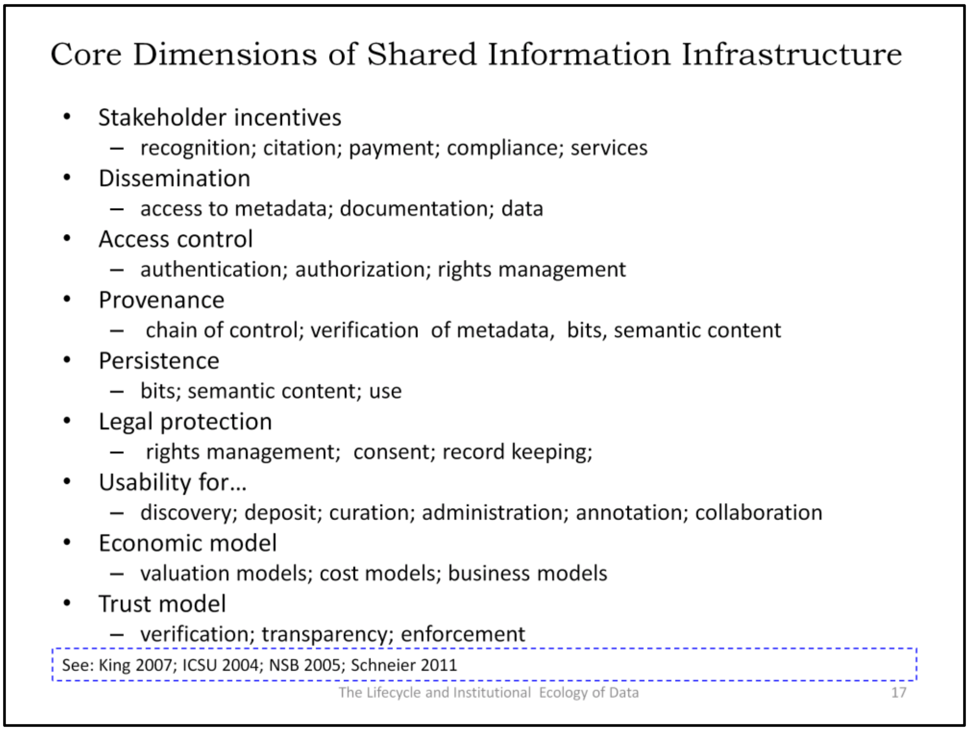
Sources/Subjects

Access Rights

Confidentiality

CNI Spring 2013 - Altman

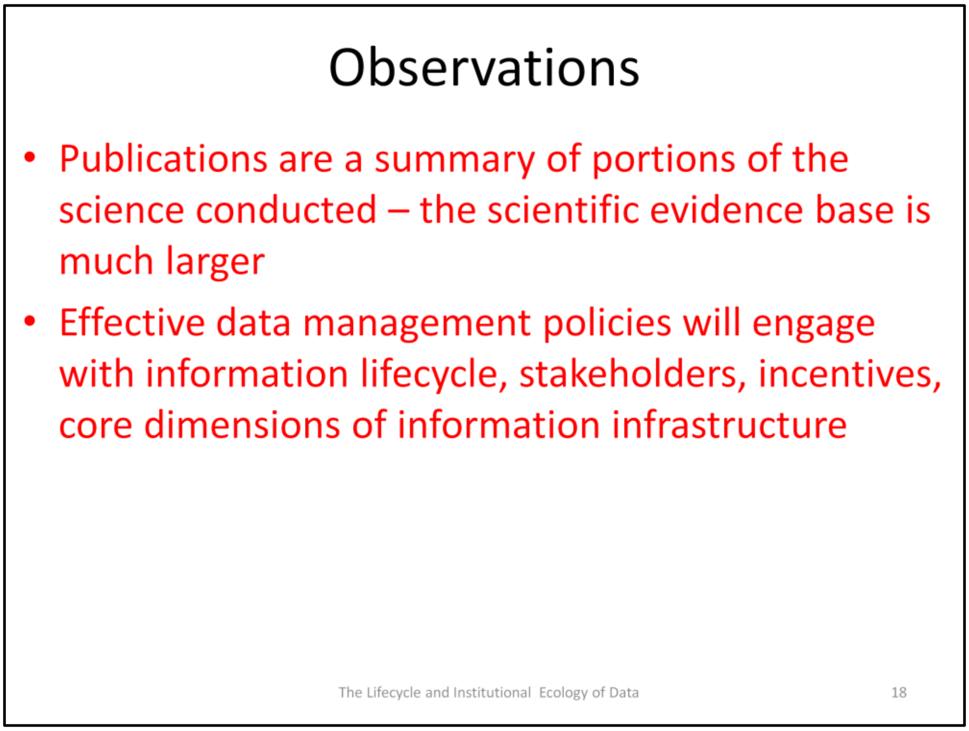
16



Core Dimensions of Shared Information Infrastructure

- Stakeholder incentives
  - recognition; citation; payment; compliance; services
- Dissemination
  - access to metadata; documentation; data
- Access control
  - authentication; authorization; rights management
- Provenance
  - chain of control; verification of metadata, bits, semantic content
- Persistence
  - bits; semantic content; use
- Legal protection
  - rights management; consent; record keeping;
- Usability for…
  - discovery; deposit; curation; administration; annotation; collaboration
- Economic model
  - valuation models; cost models; business models
- Trust model
  - verification; transparency; enforcement

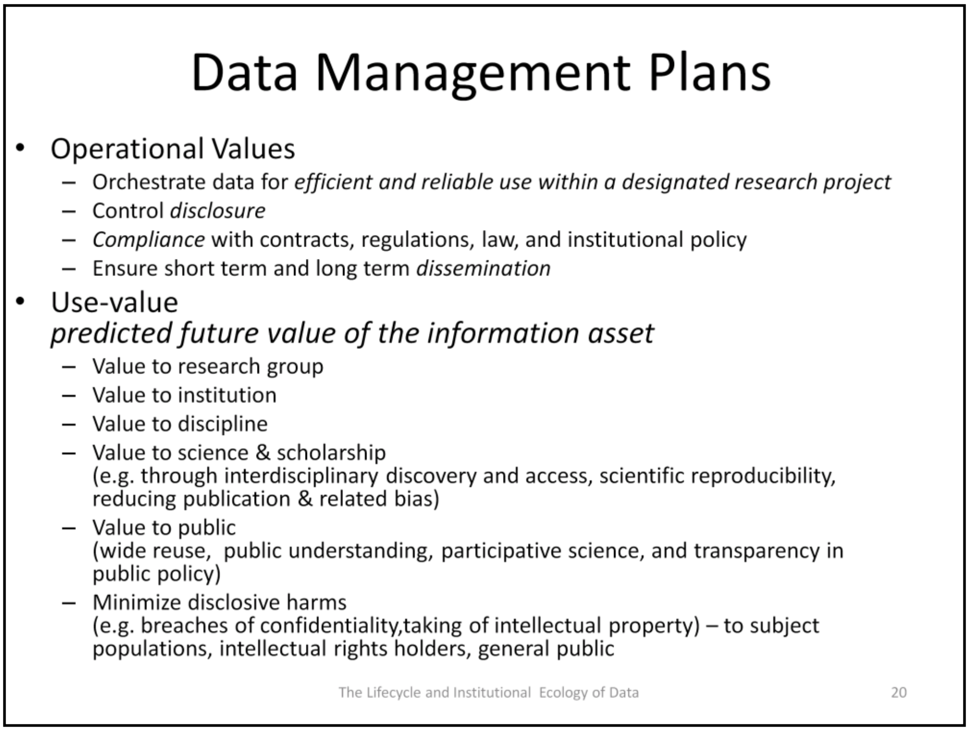
See: King 2007; ICSU 2004; NSB 2005; Schneier 2011



Observations

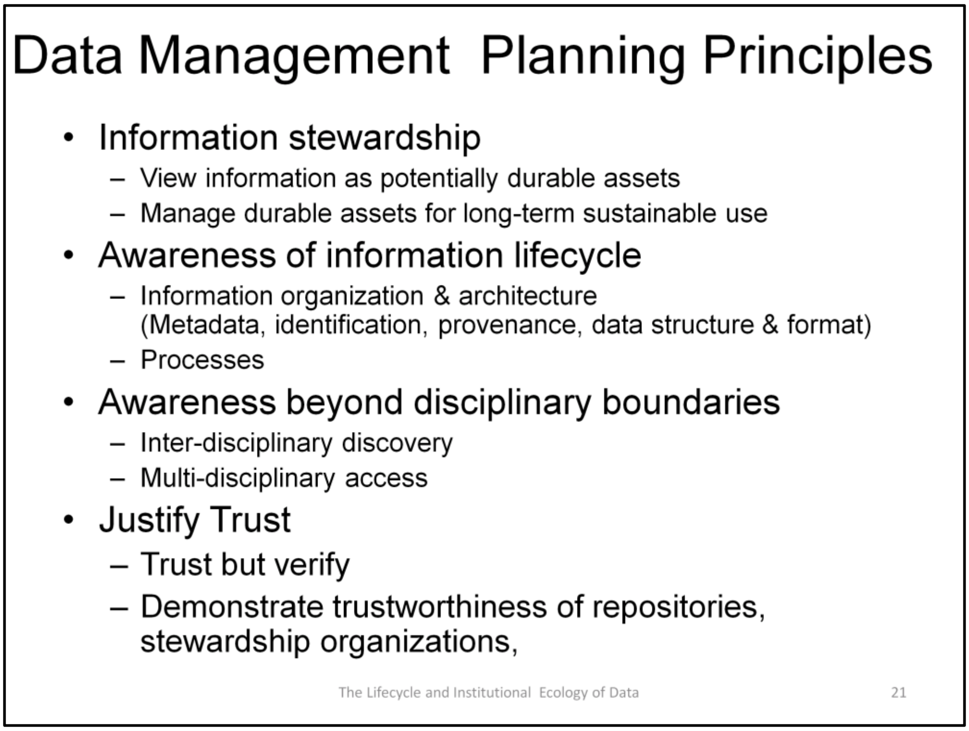
- Publications are a summary of portions of the science conducted – the scientific evidence base is much larger
- Effective data management policies will engage with information lifecycle, stakeholders, incentives, core dimensions of information infrastructure

Potential Leverage Points



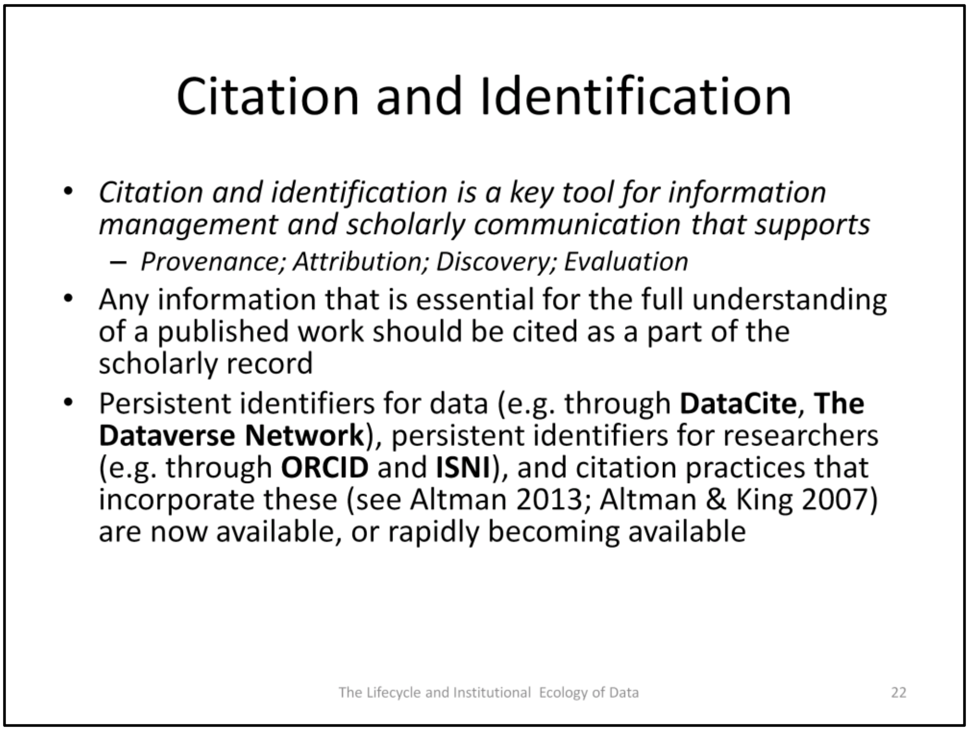
Data Management Plans

- Operational Values
  - Orchestrate data for *efficient and reliable use within a designated research project*
  - Control *disclosure*
  - *Compliance* with contracts, regulations, law, and institutional policy
  - Ensure short term and long term *dissemination*
- Use-value
  predicted future value of the information asset
  - Value to research group
  - Value to institution
  - Value to discipline
  - Value to science & scholarship
    (e.g. through interdisciplinary discovery and access, scientific reproducibility, reducing publication & related bias)
  - Value to public
    (wide reuse, public understanding, participative science, and transparency in public policy)
  - Minimize disclosive harms
    (e.g. breaches of confidentiality,taking of intellectual property) – to subject populations, intellectual rights holders, general public



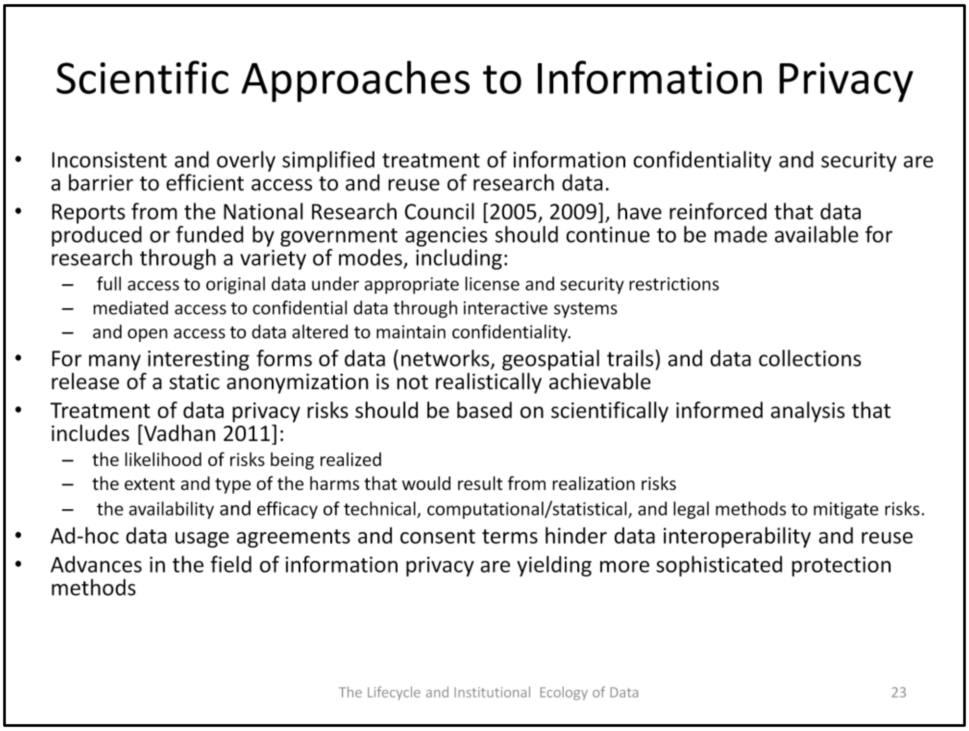
Data Management Planning Principles

- Information stewardship
  - View information as potentially durable assets
  - Manage durable assets for long-term sustainable use
- Awareness of information lifecycle
  - Information organization & architecture
    (Metadata, identification, provenance, data structure & format)
  - Processes
- Awareness beyond disciplinary boundaries
  - Inter-disciplinary discovery
  - Multi-disciplinary access
- Justify Trust
  - Trust but verify
  - Demonstrate trustworthiness of repositories, stewardship organizations,



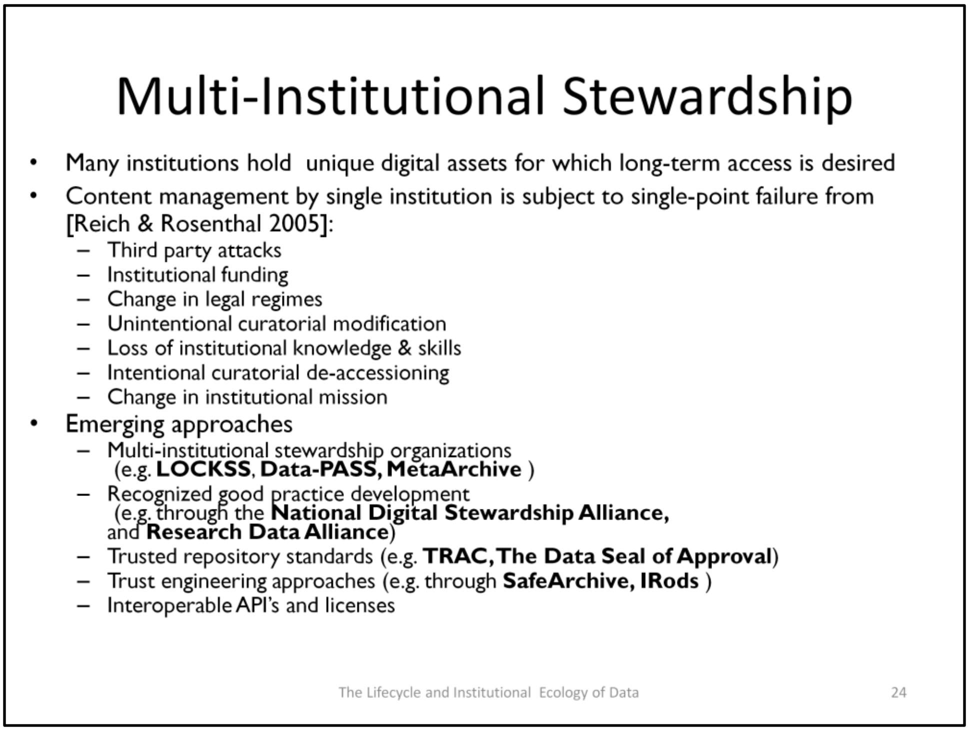
Citation and Identification

- *Citation and identification is a key tool for information management and scholarly communication that supports*
  - *Provenance; Attribution; Discovery; Evaluation*
- Any information that is essential for the full understanding of a published work should be cited as a part of the scholarly record
- Persistent identifiers for data (e.g. through **DataCite**, **The Dataverse Network**), persistent identifiers for researchers (e.g. through **ORCID** and **ISNI**), and citation practices that incorporate these (see Altman 2013; Altman & King 2007) are now available, or rapidly becoming available



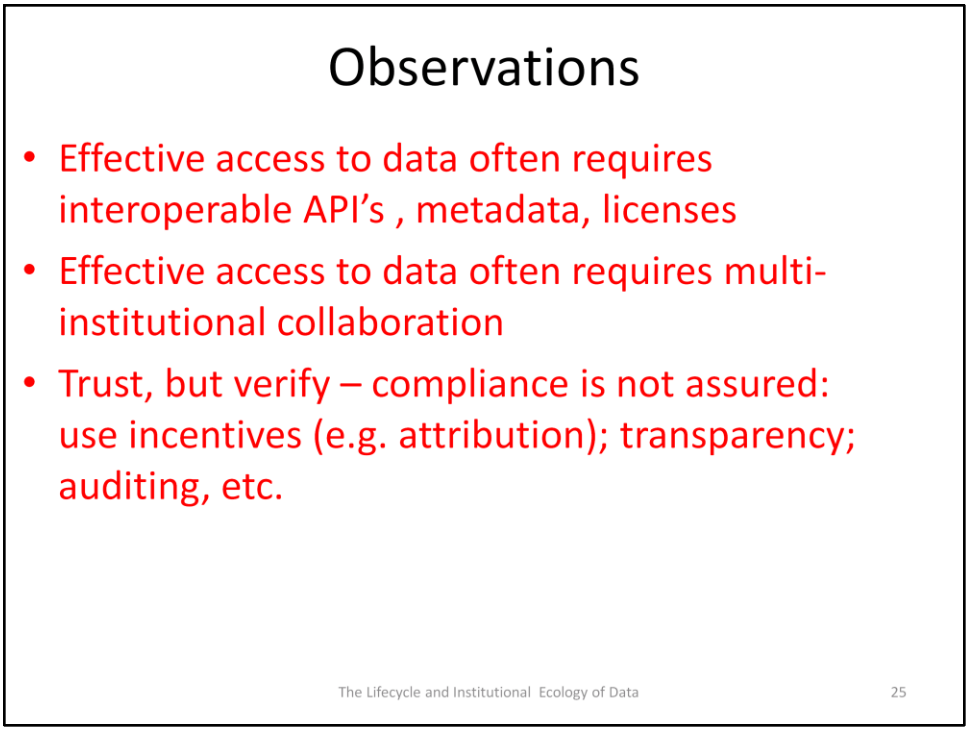
Scientific Approaches to Information Privacy

- Inconsistent and overly simplified treatment of information confidentiality and security are a barrier to efficient access to and reuse of research data.
- Reports from the National Research Council [2005, 2009], have reinforced that data produced or funded by government agencies should continue to be made available for research through a variety of modes, including:
  - full access to original data under appropriate license and security restrictions
  - mediated access to confidential data through interactive systems
  - and open access to data altered to maintain confidentiality.
- For many interesting forms of data (networks, geospatial trails) and data collections release of a static anonymization is not realistically achievable
- Treatment of data privacy risks should be based on scientifically informed analysis that includes [Vadhan 2011]:
  - the likelihood of risks being realized
  - the extent and type of the harms that would result from realization risks
  - the availability and efficacy of technical, computational/statistical, and legal methods to mitigate risks.
- Ad-hoc data usage agreements and consent terms hinder data interoperability and reuse
- Advances in the field of information privacy are yielding more sophisticated protection methods



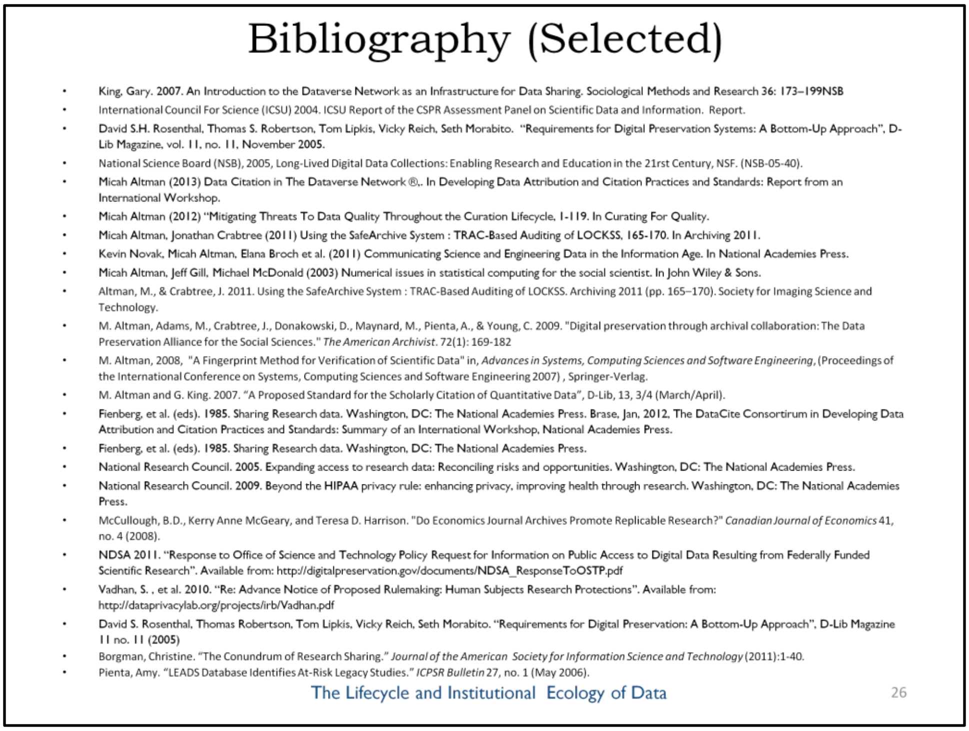
Multi-Institutional Stewardship

- Many institutions hold unique digital assets for which long-term access is desired
- Content management by single institution is subject to single-point failure from [Reich & Rosenthal 2005]:
  - Third party attacks
  - Institutional funding
  - Change in legal regimes
  - Unintentional curatorial modification
  - Loss of institutional knowledge & skills
  - Intentional curatorial de-accessioning
  - Change in institutional mission
- Emerging approaches
  - Multi-institutional stewardship organizations (e.g. **LOCKSS**, **Data-PASS, MetaArchive**)
  - Recognized good practice development (e.g. through the **National Digital Stewardship Alliance,** and **Research Data Alliance**)
  - Trusted repository standards (e.g. **TRAC, The Data Seal of Approval**)
  - Trust engineering approaches (e.g. through **SafeArchive, IRods**)
  - Interoperable API's and licenses



Observations

- Effective access to data often requires interoperable API’s , metadata, licenses
- Effective access to data often requires multi-institutional collaboration
- Trust, but verify – compliance is not assured: use incentives (e.g. attribution); transparency; auditing, etc.



Bibliography (Selected)

- King, Gary. 2007. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. Sociological Methods and Research 36: 173–199NSB
- International Council For Science (ICSU) 2004. ICSU Report of the CSPR Assessment Panel on Scientific Data and Information. Report.
- David S.H. Rosenthal, Thomas S. Robertson, Tom Lipkis, Vicky Reich, Seth Morabito. "Requirements for Digital Preservation Systems: A Bottom-Up Approach", D-Lib Magazine, vol. 11, no. 11, November 2005.
- National Science Board (NSB), 2005, Long-Lived Digital Data Collections: Enabling Research and Education in the 21rst Century, NSF. (NSB-05-40).
- Micah Altman (2013) Data Citation in The Dataverse Network ®,. In Developing Data Attribution and Citation Practices and Standards: Report from an International Workshop.
- Micah Altman (2012) "Mitigating Threats To Data Quality Throughout the Curation Lifecycle, 1-119. In Curating For Quality.
- Micah Altman, Jonathan Crabtree (2011) Using the SafeArchive System : TRAC-Based Auditing of LOCKSS, 165-170. In Archiving 2011.
- Kevin Novak, Micah Altman, Elana Broch et al. (2011) Communicating Science and Engineering Data in the Information Age. In National Academies Press.
- Micah Altman, Jeff Gill, Michael McDonald (2003) Numerical issues in statistical computing for the social scientist. In John Wiley & Sons.
- Altman, M., & Crabtree, J. 2011. Using the SafeArchive System : TRAC-Based Auditing of LOCKSS. Archiving 2011 (pp. 165–170). Society for Imaging Science and Technology.
- M. Altman, Adams, M., Crabtree, J., Donakowski, D., Maynard, M., Pienta, A., & Young, C. 2009. "Digital preservation through archival collaboration: The Data Preservation Alliance for the Social Sciences." *The American Archivist*. 72(1): 169-182
- M. Altman, 2008, "A Fingerprint Method for Verification of Scientific Data" in, *Advances in Systems, Computing Sciences and Software Engineering*, (Proceedings of the International Conference on Systems, Computing Sciences and Software Engineering 2007) , Springer-Verlag.
- M. Altman and G. King. 2007. "A Proposed Standard for the Scholarly Citation of Quantitative Data", D-Lib, 13, 3/4 (March/April).
- Fienberg, et al. (eds). 1985. Sharing Research data. Washington, DC: The National Academies Press. Brase, Jan, 2012, The DataCite Consortirum in Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop, National Academies Press.
- Fienberg, et al. (eds). 1985. Sharing Research data. Washington, DC: The National Academies Press.
- National Research Council. 2005. Expanding access to research data: Reconciling risks and opportunities. Washington, DC: The National Academies Press.
- National Research Council. 2009. Beyond the HIPAA privacy rule: enhancing privacy, improving health through research. Washington, DC: The National Academies Press.
- McCullough, B.D., Kerry Anne McGeary, and Teresa D. Harrison. "Do Economics Journal Archives Promote Replicable Research?" *Canadian Journal of Economics* 41, no. 4 (2008).
- NDSA 2011. "Response to Office of Science and Technology Policy Request for Information on Public Access to Digital Data Resulting from Federally Funded Scientific Research". Available from: http://digitalpreservation.gov/documents/NDSA_ResponseToOSTP.pdf
- Vadhan, S. , et al. 2010. "Re: Advance Notice of Proposed Rulemaking: Human Subjects Research Protections". Available from: http://dataprivacylab.org/projects/irb/Vadhan.pdf
- David S. Rosenthal, Thomas Robertson, Tom Lipkis, Vicky Reich, Seth Morabito. "Requirements for Digital Preservation: A Bottom-Up Approach", D-Lib Magazine 11 no. 11 (2005)
- Borgman, Christine. "The Conundrum of Research Sharing." *Journal of the American Society for Information Science and Technology* (2011):1-40.
- Pienta, Amy. "LEADS Database Identifies At-Risk Legacy Studies." *ICPSR Bulletin* 27, no. 1 (May 2006).

Questions?

E-mail: escience@mit.edu

Twitter: @drmaltman