Title
Sharing, Reusing, and Repurposing Data

Author
Borgman, Christine L.

Publication Date
2013-05-21

The Conundrum of Sharing Research Data

*If the rewards of the data deluge are to be reaped, then researchers who produce those data must share them, and do so in such a way that the data are interpretable and reusable by others.**

*Borgman, C.L. (2012). The Conundrum of Sharing Research Data. JASIST, 63(6):1059–1078

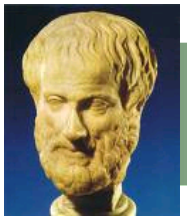
http://www.tzanis.org/tzanisblog/archives/images/push-pull-thumb.jpg

Overview

- Paradigm shift
- Arguments for sharing data
- Science friction, data friction
- Success factors for reusing and repurposing data

http://inventionmachine.com/the-Invention-Machine-Blog/bid/51703/Three-Key-Challenges-to-Entering-New-Markets

New problem solving methods



Empirical

Theory

Simulation

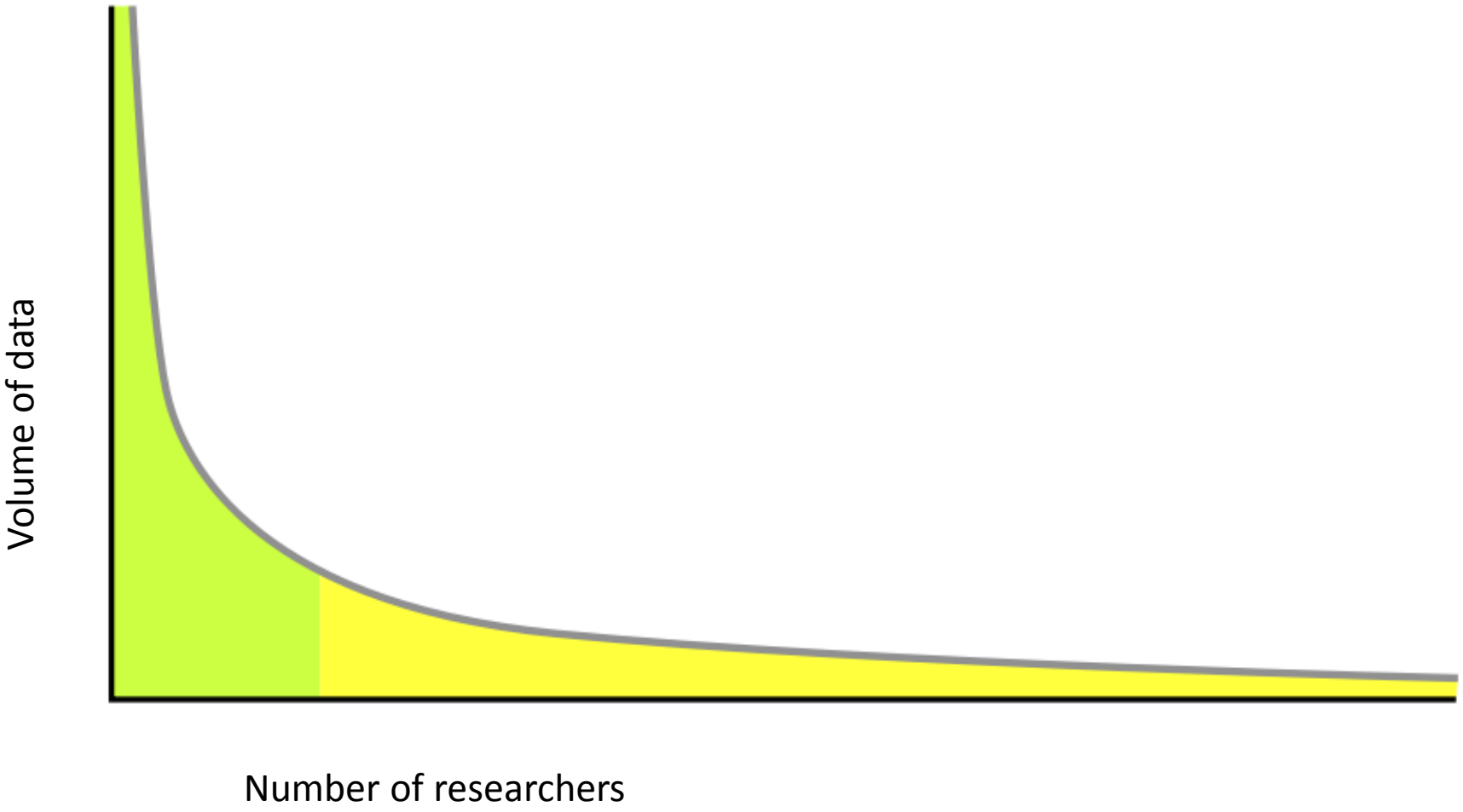
Data

"Applied computer science is now playing the role that mathematics did from the 17th through the 20th centuries: providing an orderly, formal framework and exploratory apparatus for other sciences"

– G. Djorgovski

Slide courtesy Ian Foster, 2009

The long tail of data



Slide: The Institute for Empowering Long Tail Research

Big Data

Little Data

No Data

No Data is the Norm

Data sharing imperatives

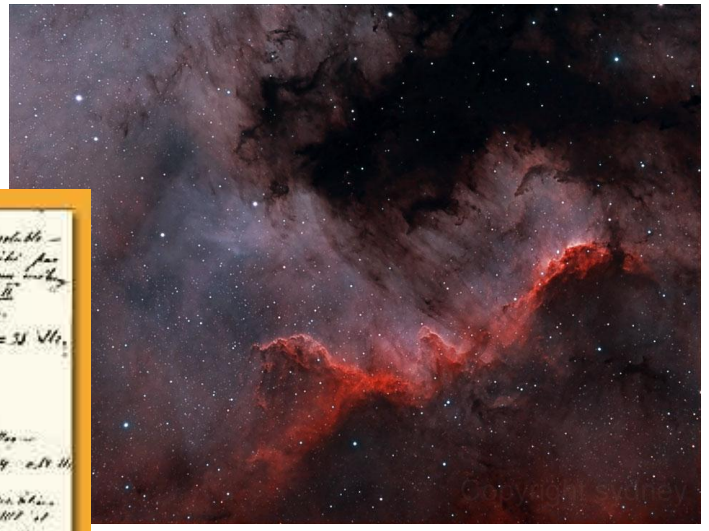
- Research Councils of the UK
  - Open access publishing requirements
  - Provisions for access to data
- Wellcome Trust
  - Open access publishing
  - Data sharing requirements
- National Science Foundation
  - Data sharing requirements
  - Data management plans
- U.S. Federal policy-2013
  - Open access to publications
  - Open access to data

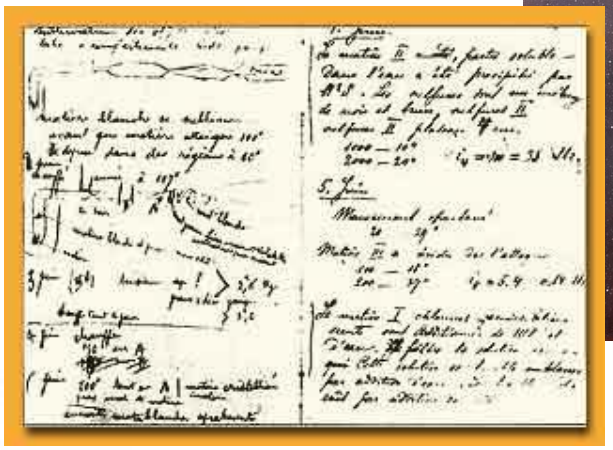
Supported by wellcome trust


Science
12 September 2008 | $10
AAA


Science
11 February 2011 | $10
scientific
data
information
climate
research
science
new
analysis
visualization
researchers
access
many
human
knowledge
example
AAAS

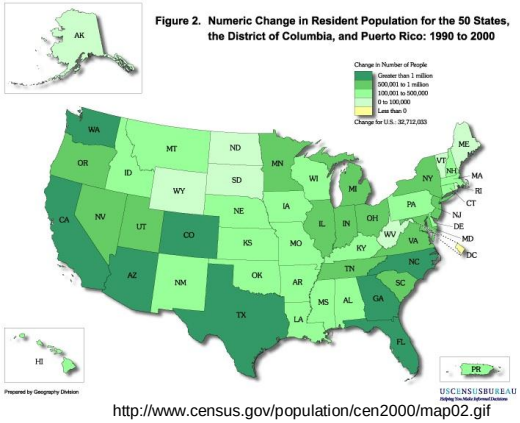
What are data?



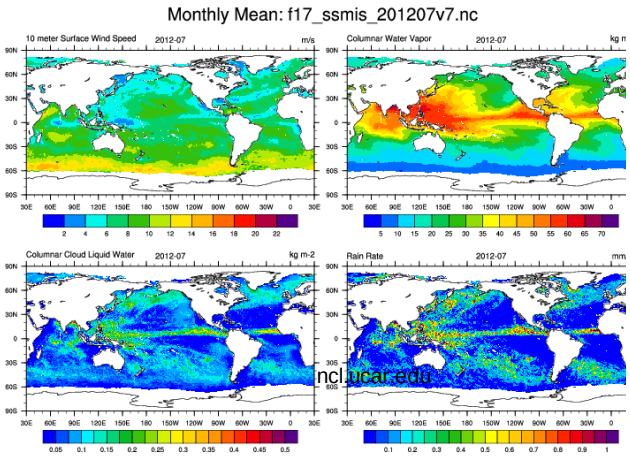


Marie Curie's notebook aip.org

hudsonalpha.org



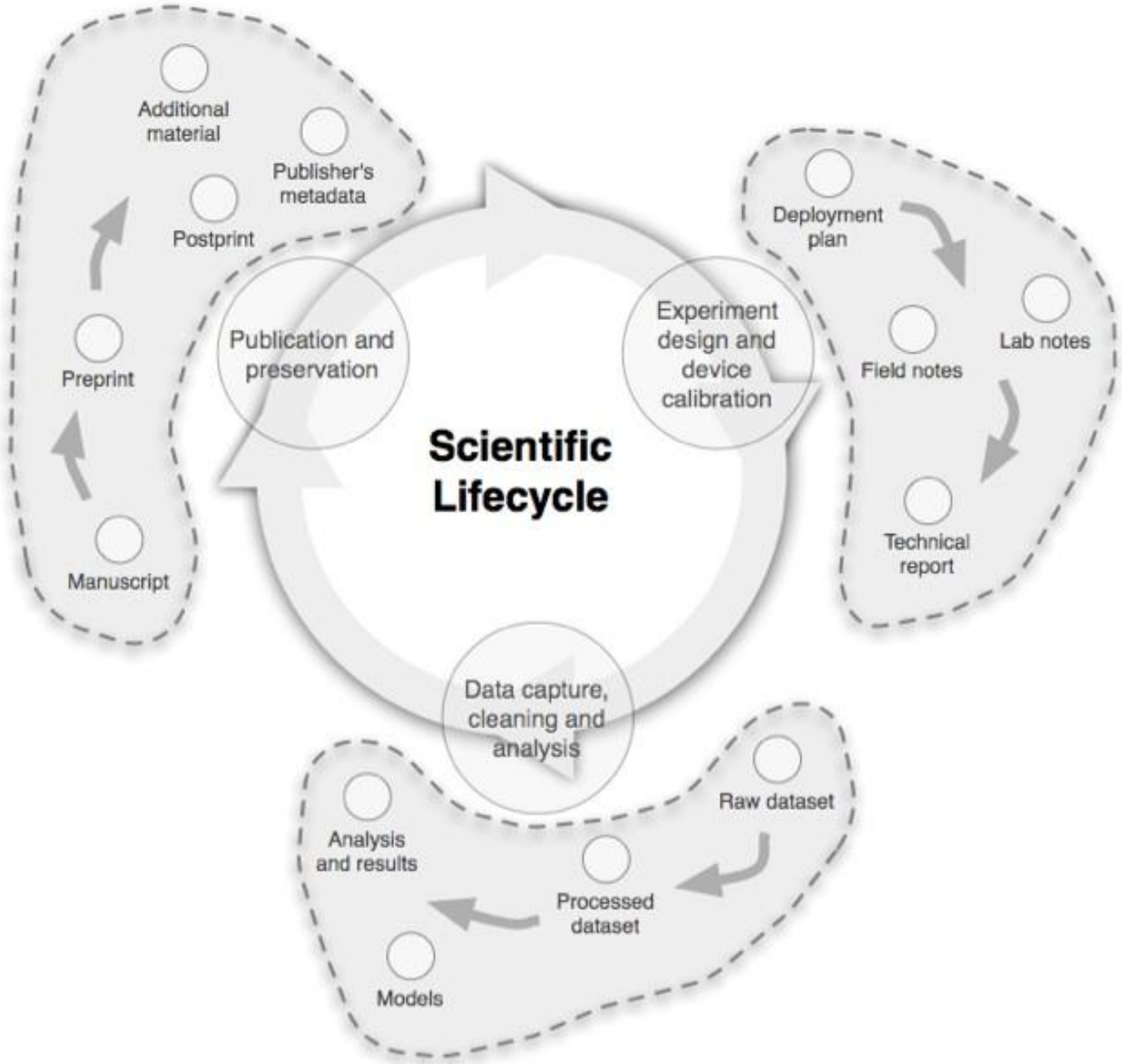
http://www.census.gov/population/cen2000/map02.gif



Date:1/2.07.75 Place:Sakaltutan
Zafor
He will grow old in his present house; new house is for sons - 5 sons. Not sure they want to live in village. He will only build another if they want him to. eS came from Germany and did the plastering. He arranged the carpentry in Kayseri. Çok para gitti. {much money went} Has a tractor.

Date:July1980 Place:Sakaltutan
Zafor:
Household now Zafor and wife; Nazif Unal and wife and youngest son, still a boy. They run two dolmuß; one with a driver from Süleymanli. Goes in and out once a day. He gets 8,000 a month. Zafor then said, keskin de©il. { not sharp - i.e.? not profitable} I said he did very well on 8,000 TL with only two journeys a day. Nazif Unal has "bought" a Durak {dolmuß stop} from Belediye and works all day in Kayseri.

http://onlineqda.hud.ac.uk/Intro_QDA/Examples_of_Qualitative_Data.php



Pepe, A., Mayernik, M. S., Borgman, C. L. & Van de Sompel, H. (2010). From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web. Journal of the American Society for Information Science and Technology, 61(3): 567–582.

Overview

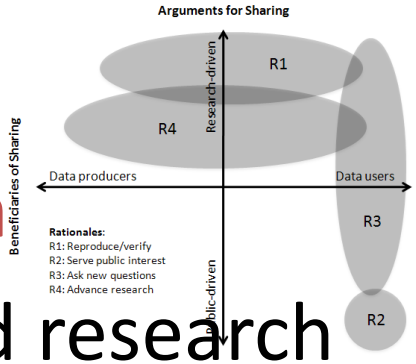
- Paradigm shift
- Arguments for sharing data
- Science friction, data friction
- Success factors for reusing and repurposing data

http://inventionmachine.com/the-Invention-Machine-Blog/bid/51703/Three-Key-Challenges-to-Entering-New-Markets

Why share research data?

Rationales

1. To reproduce or to verify research
2. To make results of publicly funded research available to the public
3. To enable others to ask new questions of extant data
4. To advance the state of research and innovation

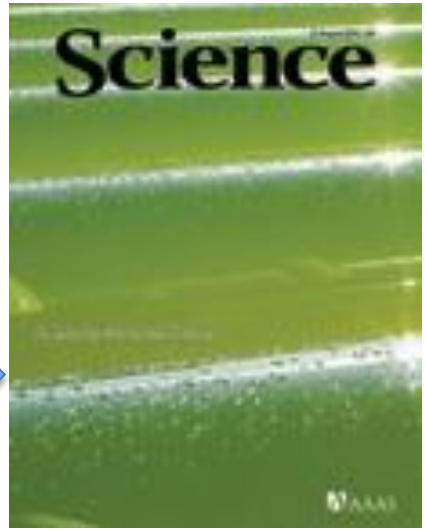


Borgman, C.L. (2012). The Conundrum of Sharing Research Data. JASIST, 63(6):1059–1078

1. Reproduce or verify research

http://chemistry.curtin.edu.au/research/index.cfm

Benzoic Acid	% yield		IR Peaks (cm^{-1})		Solid (C) or Oil (O) Product	Mp (^{0}C)
	Gross	Recrystallization	N-H	C=O		
Sodium benzoate		2.58	3327	1638	White C	79-89
Sodium benzoate			3337	1640&1600	O	
Sodium benzoate			3326	1642&1601	O	
Sodium benzoate	37.8		3274	1640	O	
p-nitro	51.84	10.59	3423	1693	Yellow C	152-157
m-nitro	37.38	5.43	3334	1694	Green C	152-157
Benzoic acid		7.44	3293	1642	White C	152-154
m-bromo		47.4	3316	1702	Green paste	
p-bromo		14.53	3344	1638	Pink C	164-166
p-chloro		29.69	3340	1638	Yellow C	
m-chloro		74.53	3410	1637	tan paste	
o-chloro		17.31	3422	1654	Tan C	
3,5-dinitro		44.53	3297	1647	Tan C	139-141
p-hydroxy		3.751	3401	1643	yellow/green C	210
p-amino		8.475	3411	1645	Dark O	
o-methoxy		42.49	3412	1646	Yellow O	



http://serc.carleton.edu/cismi/broadaccess/groupwork.html

Scientific Gold Standard

REPLICATION—THE CONFIRMATION OF RESULTS AND CONCLUSIONS FROM ONE STUDY obtained independently in another—is considered the scientific gold standard.

Jasny, B. R., Chin, G., Chong, L. & Vignieri, S. (2011). Again, and again, and again. Science, 334(6060): 1225.

Victoria Stodden, Columbia

Reproducibility?

- Deductive sciences
  - Check the proof
- Experimental sciences
  - Redo the field work
- Computational sciences
  - Start with the dataset
  - Reconstruct workflow

Analytic validity	Do different labs, techniques, and platforms measure the same thing?
Repeatability	Can other scientists access the data and protocols, repeat the analyses, and get the same results?
Replication	Do many different data sets and their combination (meta-analysis) get consistent results?
External validation	Do different data sets by different teams, preferably prospectively and with large-scale evidence, get consistent results?
Clinical validity	Does the discovered information predict clinical outcomes?
Clinical utility	Does the use of the discovered information improve clinical outcomes?

Science AAAS

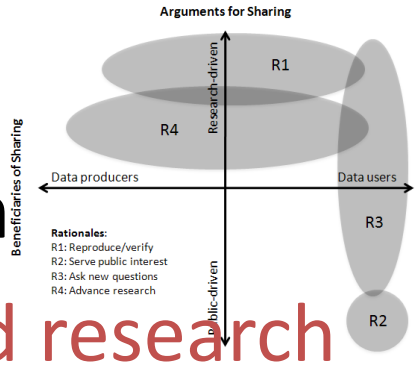
Published by AAAS

J P A Ioannidis, M J Khoury Science 2011;334:1230-1232

Why share research data?

Rationales

1. To reproduce or to verify research
2. To make results of publicly funded research available to the public
3. To enable others to ask new questions of extant data
4. To advance the state of research and innovation



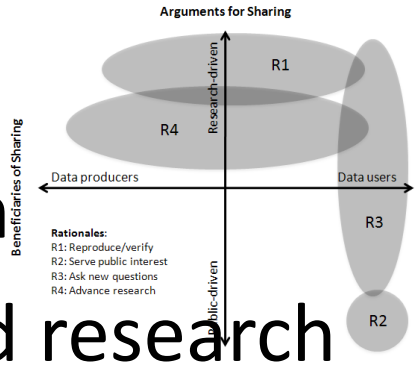
Borgman, C. L. (2012, forthcoming). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*. Figure by Jillian C. Wallis, UCLA

2. Public monies serve the public good

Why share research data?

Rationales

1. To reproduce or to verify research
2. To make results of publicly funded research available to the public
3. To enable others to ask new questions of extant data
4. To advance the state of research and innovation

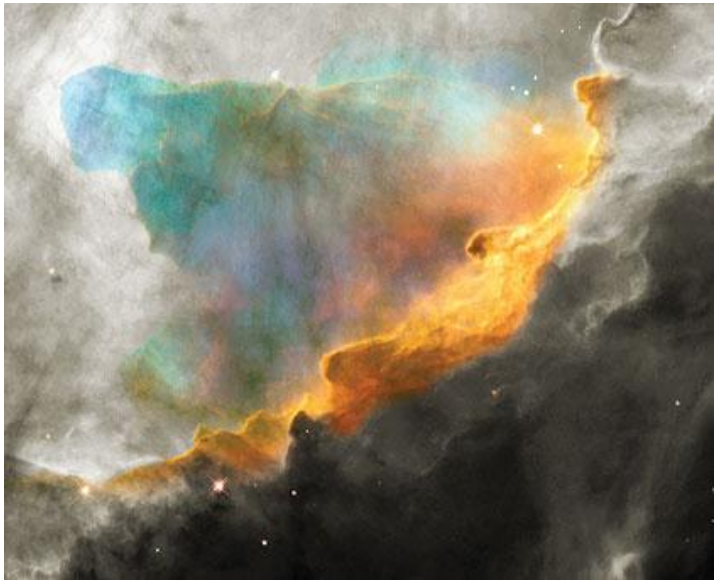


Borgman, C. L. (2012, forthcoming). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*. Figure by Jillian C. Wallis, UCLA

3. Others can ask new questions

data

discovery



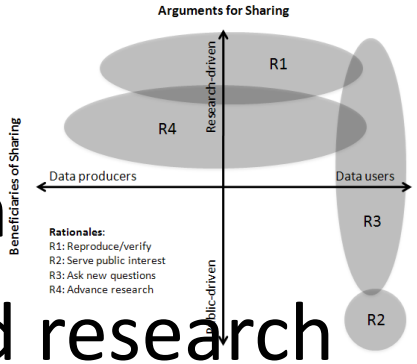
http://annualreport.ucdavis.edu/2008/images/photos/discovery.jpg

http://digitalassetmanagement.org.uk/2010/02/01/the-winds-of-change-are-blowing-in-the-clouds-favor/

Why share research data?

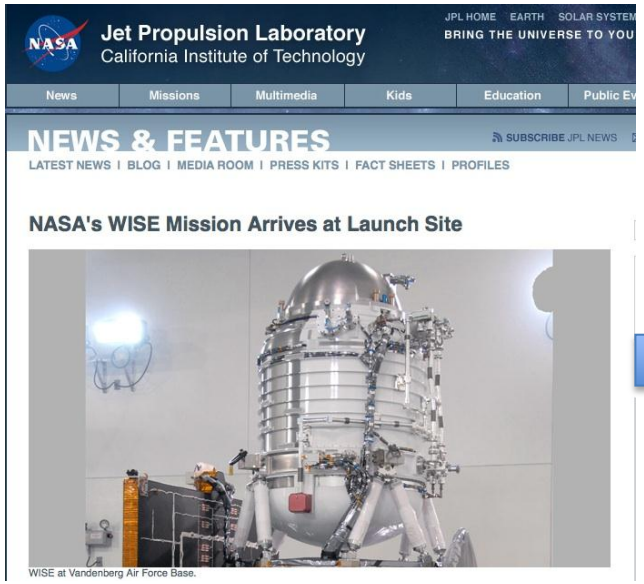
Rationales

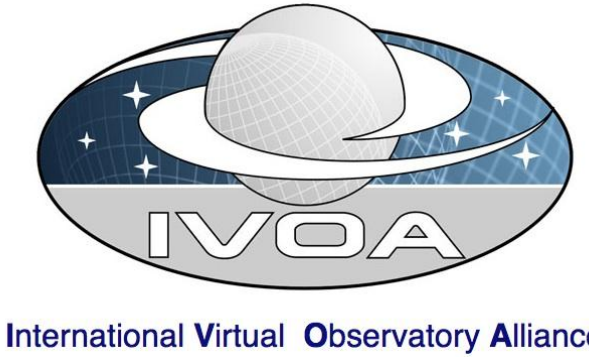
1. To reproduce or to verify research
2. To make results of publicly funded research available to the public
3. To enable others to ask new questions of extant data
4. To advance the state of research and innovation



Borgman, C. L. (2012, forthcoming). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*. Figure by Jillian C. Wallis, UCLA

4. Data curation advances research

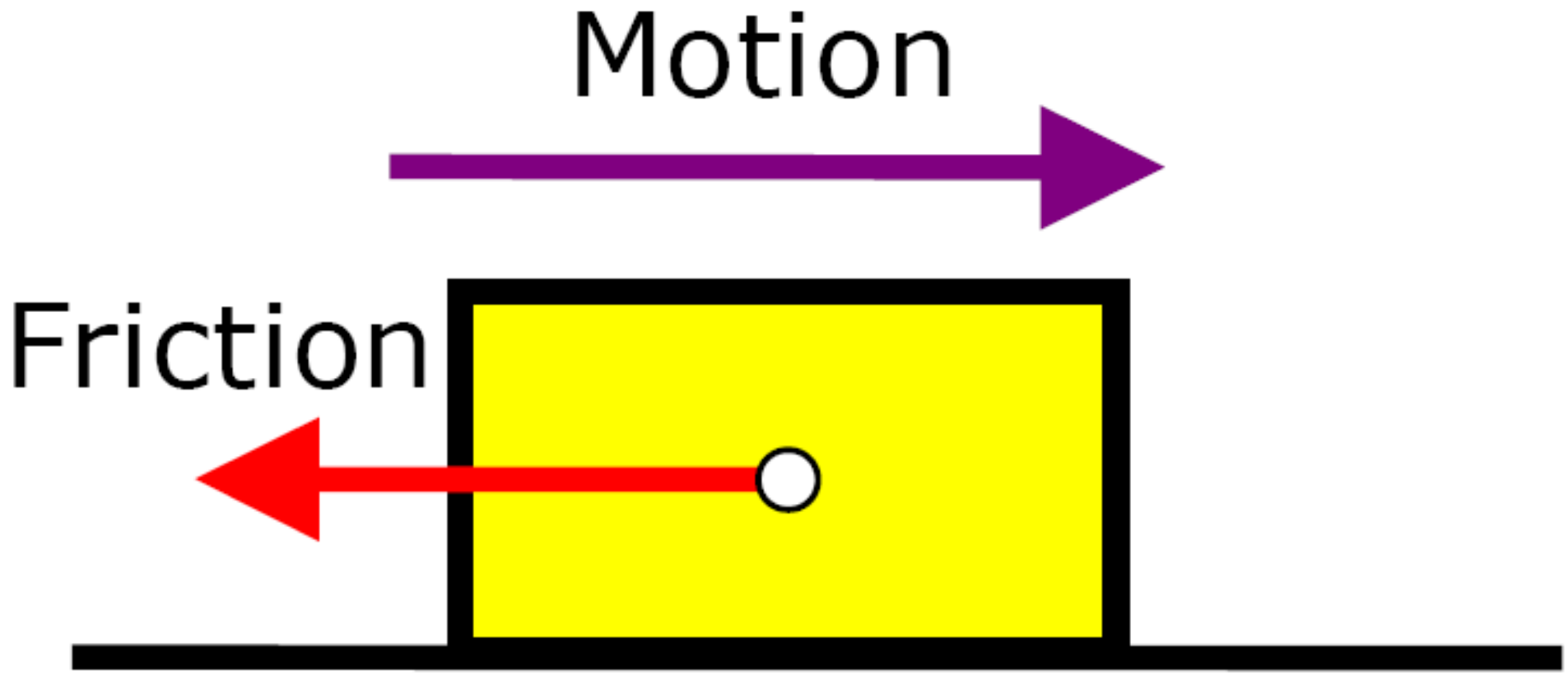




Overview

- Paradigm shift
- Arguments for sharing data
- Science friction, data friction
- Success factors for reusing and repurposing data

http://inventionmachine.com/the-Invention-Machine-Blog/bid/51703/Three-Key-Challenges-to-Entering-New-Markets



http://www.stmary.ws/highschool/physics/home/notes/dynamics/friction/imgE2.gif

Science friction, data friction*

- Data are unruly objects
- Data do not stand alone
- Data reuse is a function of distance from origin
- Intractable problems

*Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science Friction: Data, Metadata, and Collaboration. *Social Studies of Science*, *41*, 667–690. doi:10.1177/0306312711413314

www.zazzle.com

Data are unruly objects*

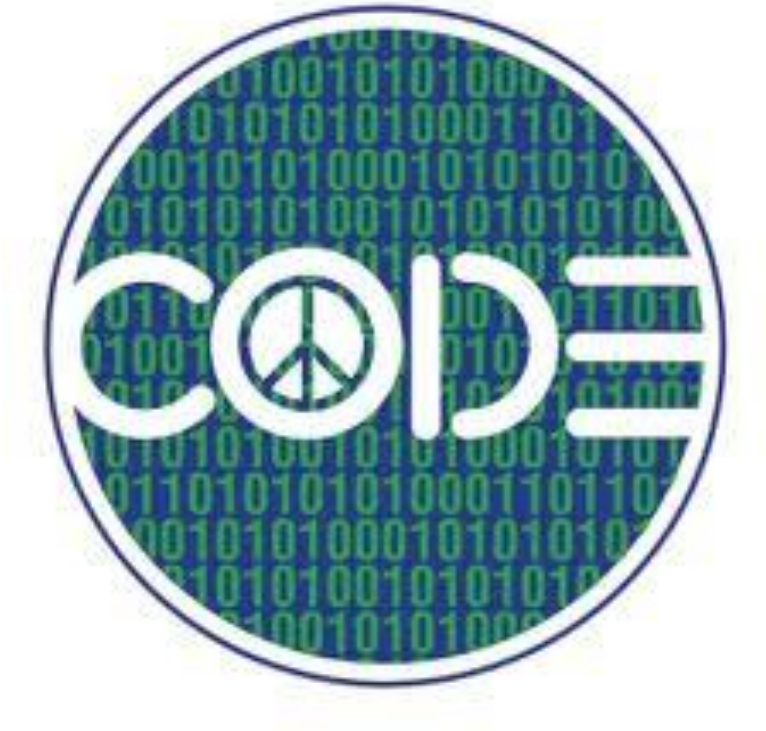
- Poorly bounded
- Malleable, mutable, mobile (Latour)
- Dynamic, evolving
- Signal to noise varies by use

www.zazzle.com

*Wynholds, L. A. (2010). Linking to Scientific Data: Identity Problems of Unruly and Poorly Bounded Digital Objects. Presented at the Digital Curation Conference, 15 June 2011. http://www.ijdc.net/index.php/ijdc/article/view/174

Data do not stand alone

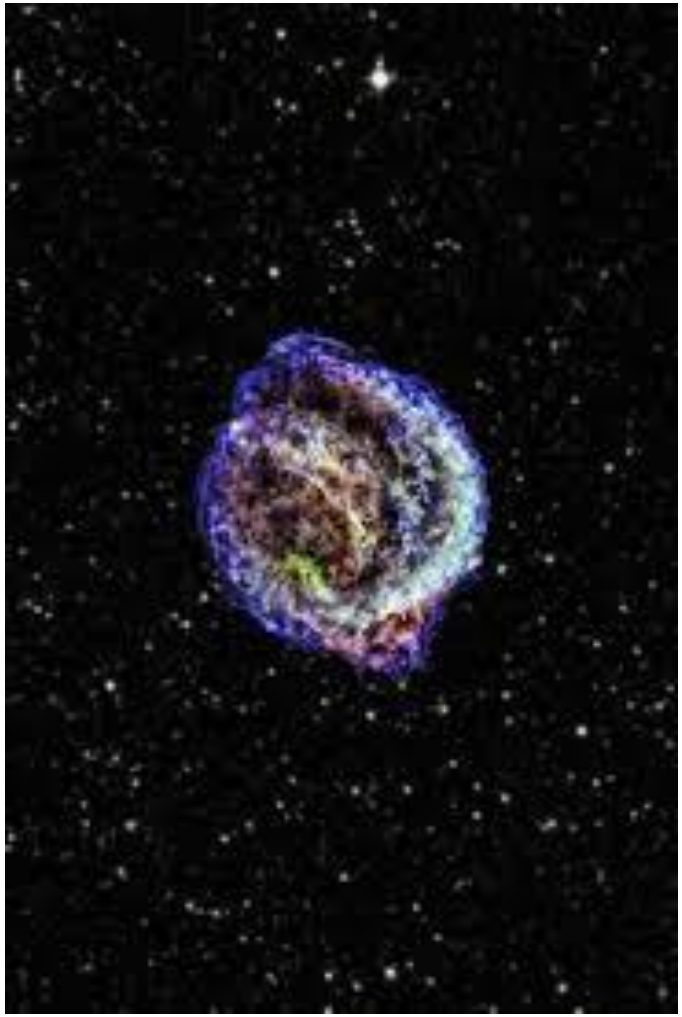
- Data are inseparable
  - Code
  - Technical standards
  - Documentation
  - Instrumentation
  - Calibration
  - Provenance
  - Workflows
  - Local practices
  - Physical samples



http://peacetour.org/sites/default/files/code4peace-logo2-v3-color-sm.jpg

Data reuse is a function of distance from origin

- Reuse by investigator
- Reuse by collaborators
- Reuse by colleagues
- Reuse by unaffiliated others
- Reuse at later times
  - Months
  - Years
  - Decades
  - Centuries



http://chandra.harvard.edu/photo/2013/kepler/kepler_525.jpg

Intractable problems

- Confidentiality
- Anonymization
- Reidentification
- Intellectual property
- Economics

http://fyi.uiowa.edu/wp-content/uploads/2011/10/utopia_in_four_movements_filmstill5_utopiasign.jpg

Overview

- Paradigm shift
- Arguments for sharing data
- Science friction, data friction
- Success factors for reusing and repurposing data

http://inventionmachine.com/the-Invention-Machine-Blog/bid/51703/Three-Key-Challenges-to-Entering-New-Markets

The Conundrum of Sharing Research Data

*If the rewards of the data deluge are to be reaped, then researchers who produce those data must share them, and do so in such a way that the data are interpretable and reusable by others.**

*Borgman, C.L. (2012). The Conundrum of Sharing Research Data. JASIST, 63(6):1059–1078

http://www.tzanis.org/tzanisblog/archives/images/push-pull-thumb.jpg

How to share data

- Curated data archive: NASA, UKDA, ICPSR…
- Author curated data archive
- University data archive: ORA
- Personal website
- ftp site
- Email on request

http://www.zippykidstore.com/

Simple Rules for the Care and Feeding of Scientific Data*

1. Good science requires good data
2. Make your science inspectable by others
3. Conduct your science with provenance in mind
4. Do not reduce your data more than necessary
5. Make your data available
6. Make your workflows available
7. Publish all software, even small scripts
8. Foster a “data community” for your community
9. Describe how you want to be acknowledged
10. Attribute the sources of data that you use

*DRAFT: Radcliffe Seminar on Data Provenance, 9-10 May 2013, A. Goodman & X-L Meng

Conclusions

- Data reuse is part of open science / open scholarship
- Data sharing is a paradigm shift
- Data are not journal articles (yet)
- Data are messy
- Data sharing is a necessary but not sufficient condition for reuse
- Data reuse depends on
  - Conditions of sharing
  - Conditions of reuse
- Data friction is part of scholarship
- Better practices in managing data will increase [illegible]ata

http://www.tzanis.org/tzanisblog/archives/images/push-pull-thumb.jpg

Acknowledgements

- National Science Foundation
  - *CENS*: Cooperative Agreement #CCR-0120778, D.L. Estrin, UCLA, PI.
  - *CENS Education Infrastructure*: #ESI- 0352572, W.A. Sandoval, PI; C.L. Borgman, co-PI.
  - *Towards a Virtual Organization for Data Cyberinfrastructure*, #OCI-0750529, C.L. Borgman, UCLA, PI; G. Bowker, Santa Clara University, Co-PI; T. Finholt, University of Michigan, Co-PI.
  - *Monitoring, Modeling & Memory: Dynamics of Data and Knowledge in Scientific Cyberinfrastructures*: #0827322, P.N. Edwards, UM, PI; Co-PIs C.L. Borgman, UCLA; G. Bowker, SCU; T. Finholt, UM; S. Jackson, UM; D. Ribes, Georgetown; S.L. Star, SCU)
  - *Data Conservancy*: OCI0830976, Sayeed Choudhury, PI, Johns Hopkins University.
  - Knowledge and Data Transfer: the Formation of a New Workforce. # 1145888. C.L. Borgman, PI; S. Traweek, Co-PI.
- Microsoft External Research: Tony Hey, Lee Dirks, Catherine van Ingen, Catherine Marshall
- Sloan Foundation: The Transformation of Knowledge, Culture, and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective. # 20113194. C.L. Borgman, PI; S. Traweek, Co-PI. Joshua Greenberg, program director
- Project website: http://knowledgeinfrastructures.gseis.ucla.edu/index.html