Knowledge Infrastructures: The Invisible Foundation of Research Data

Or, How Infrastructure Connects and Disconnects Research Communities

Christine L. Borgman

Distinguished Research Professor

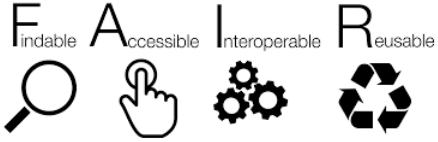
University of California, Los Angeles

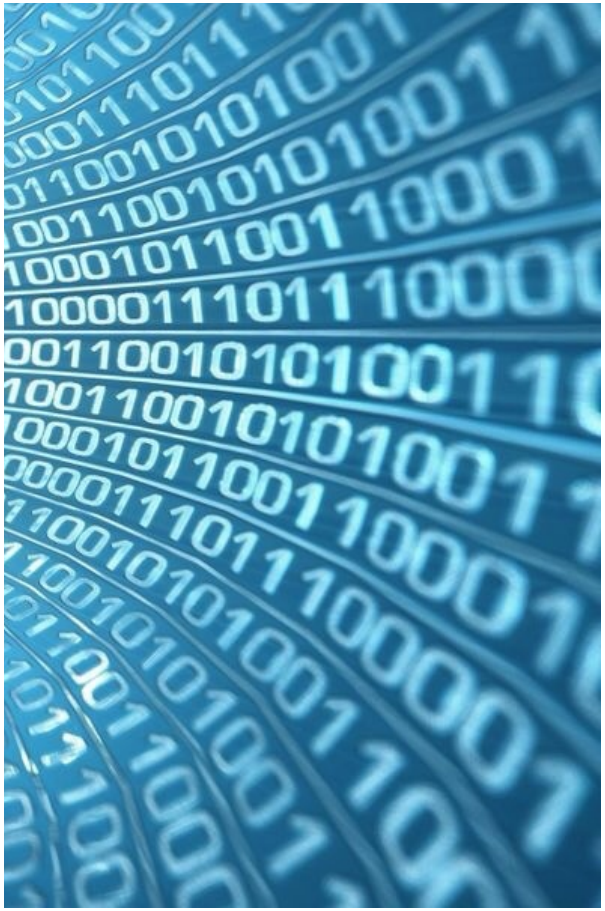
http://christineborgman.info

First Conference on Research Data Infrastructure
Karlsruhe, Germany
Keynote Presentation
12 September 2023

Why Research Data Infrastructure?

- Research data
  - are valuable entities worthy of stewardship
  - are useful to others
  - will be reused
  - should be findable, accessible, interoperable, and reusable





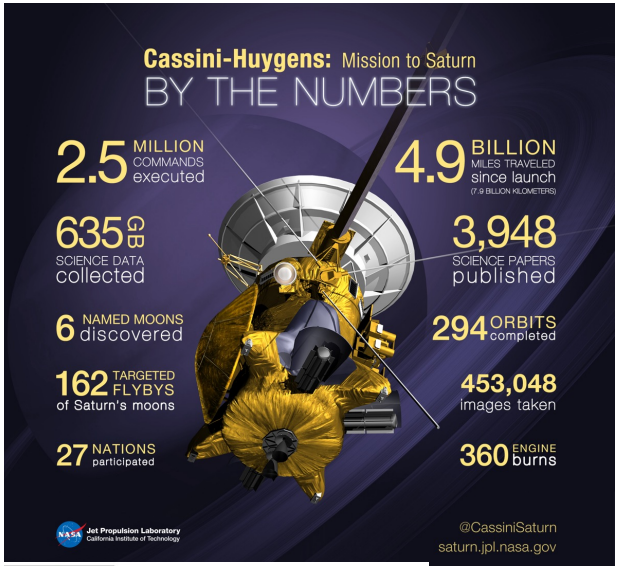
Research data infrastructure: Stakeholders

- Research funding agencies
- Individual scientists and scholars
- Academic institutions
  - Academic leadership
  - Research computing
  - University libraries
  - Schools and departments
- Students and teachers
- General public

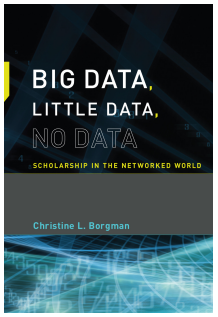
Photo by Mihai Surdu on Unsplash

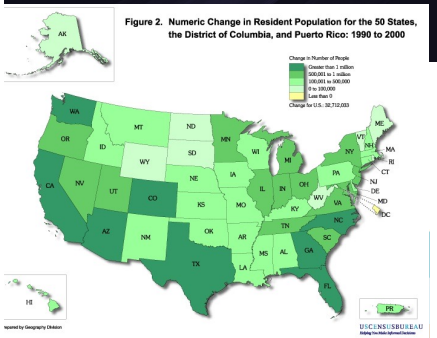
Borgman, C. L., & Bourne, P. E. (2022). Why It Takes a Village to Manage and Share Data. *Harvard Data Science Review*, *4*(3).

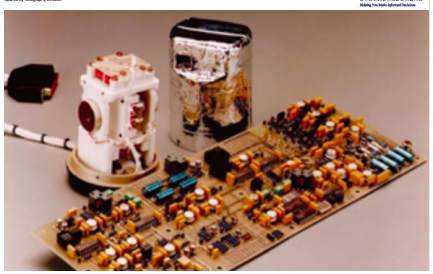
Borgman, C. L., & Brand, A. (2022). Data blind: Universities lag in capturing and exploiting data. *Science*, *378*(6626), 1278–1281.

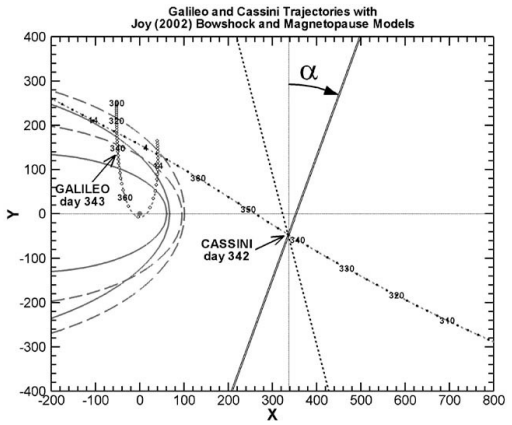


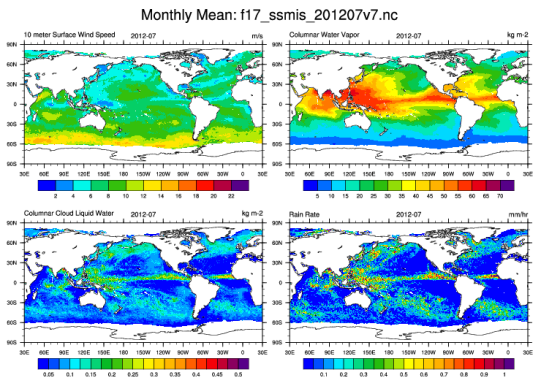
Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship.











Kivelson, M. G., & Southwood, D. J. (2003). First evidence of IMF control of Jovian magnetospheric boundary locations: Cassini and Galileo magnetic field measurements compared. *Planetary and Space Science*, *51*(13), 891–898. https://doi.org/10.1016/S0032-0633(03)00075-8

Infrastructure

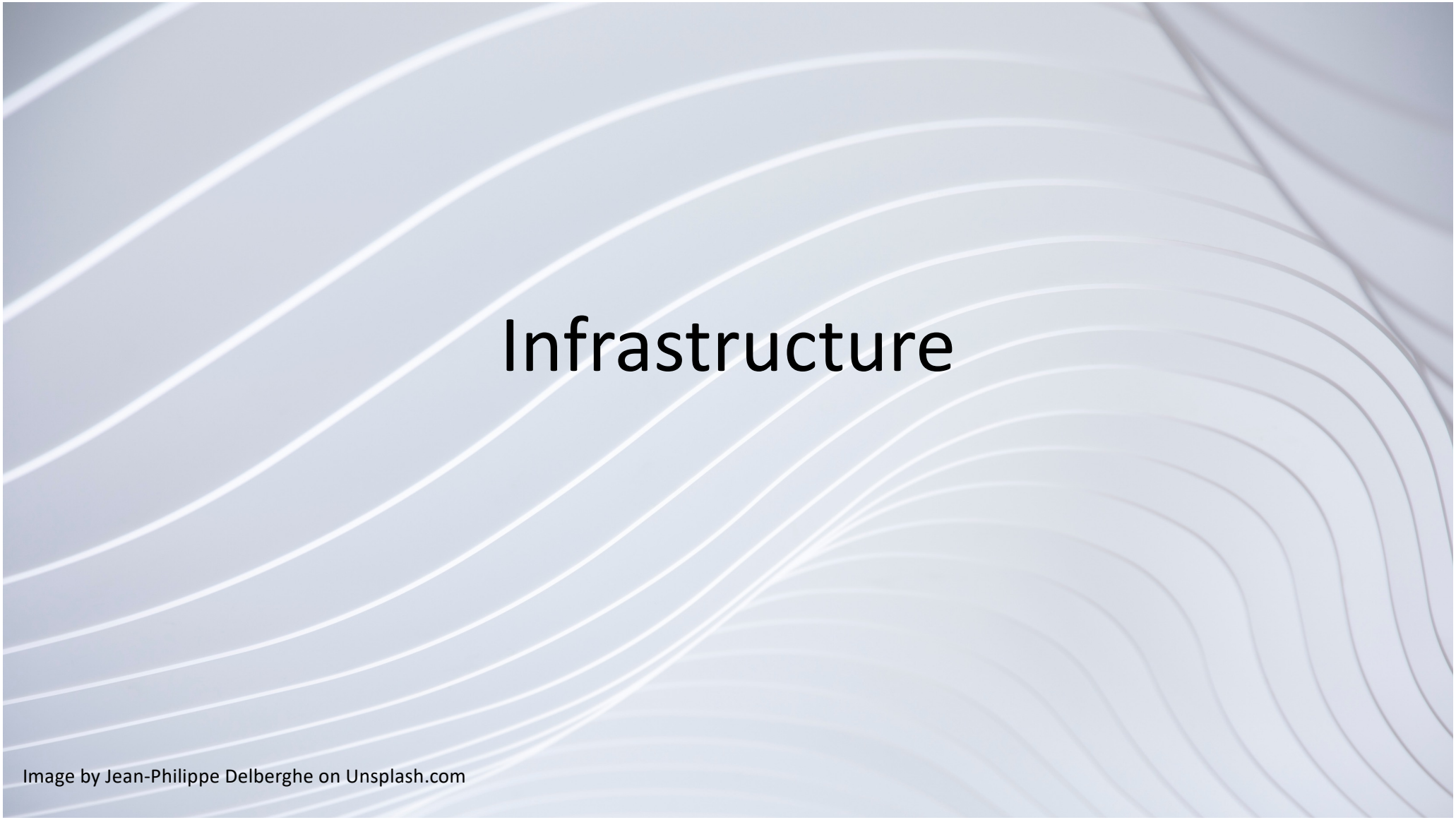


Image by Jean-Philippe Delberghe on Unsplash.com

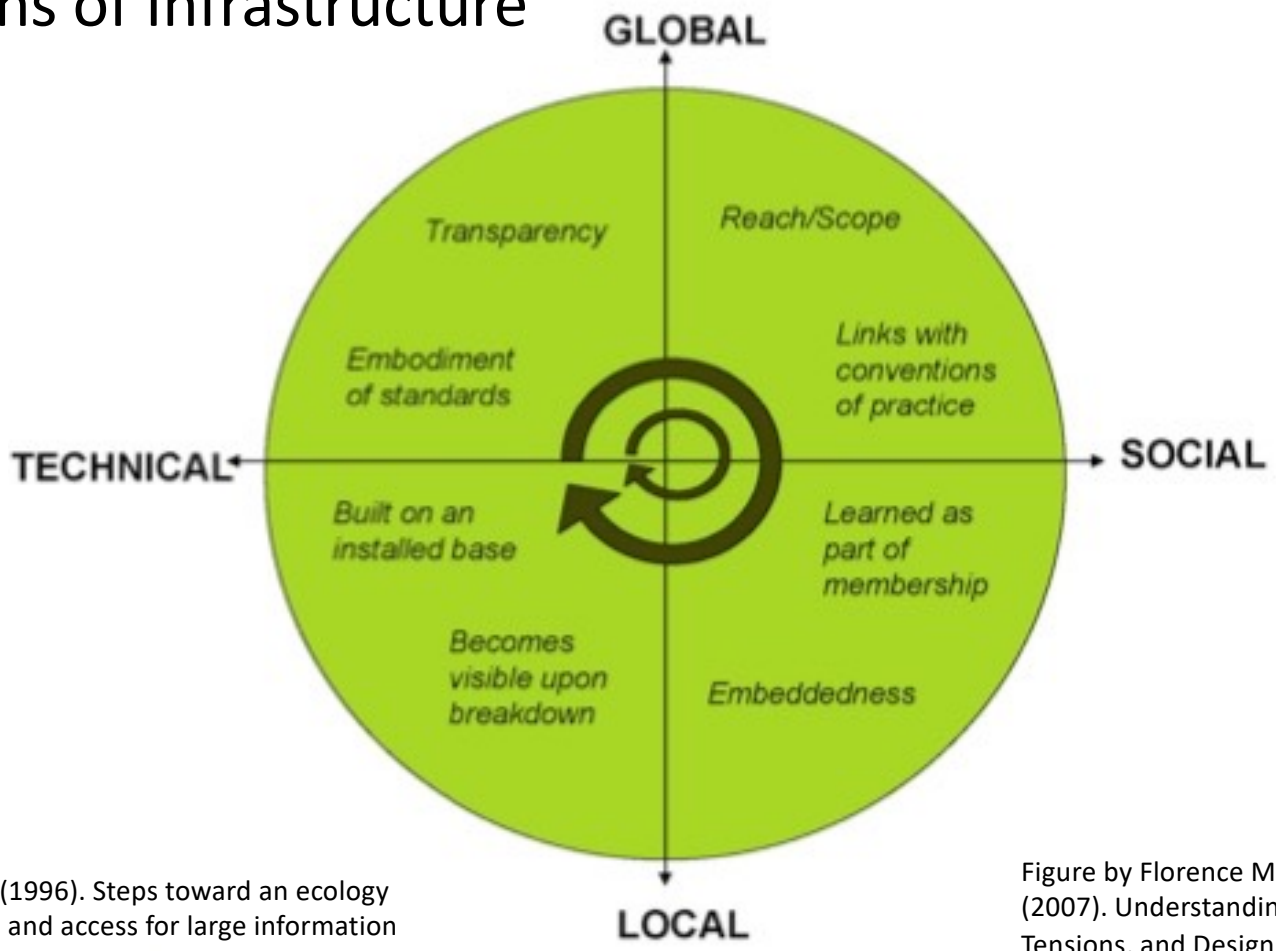
Knowledge infrastructures

Robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds (Edwards, 2010)

- Technical infrastructures
- Scholarly practices
- Policy frameworks
- Governance models

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. MIT Press.

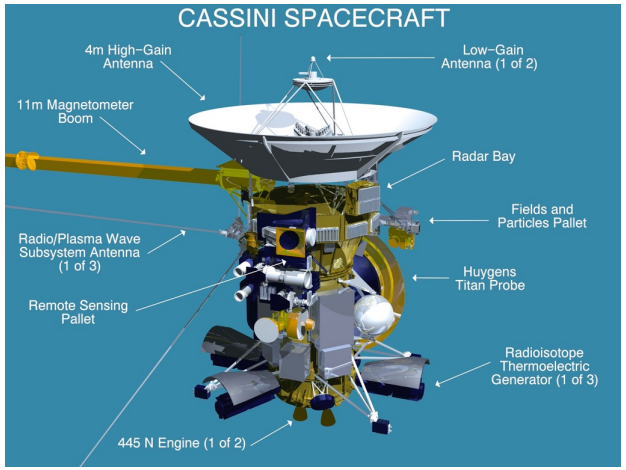
Dimensions of Infrastructure



Star, S. L. & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. Information Systems Research, 7(1): 111-134.

Figure by Florence Millerand, from: Edwards, et al (2007). Understanding Infrastructure: Dynamics, Tensions, and Design. National Science Foundation: University of Michigan. NSF Grant 0630263.

Infrastructure: Global and Technical

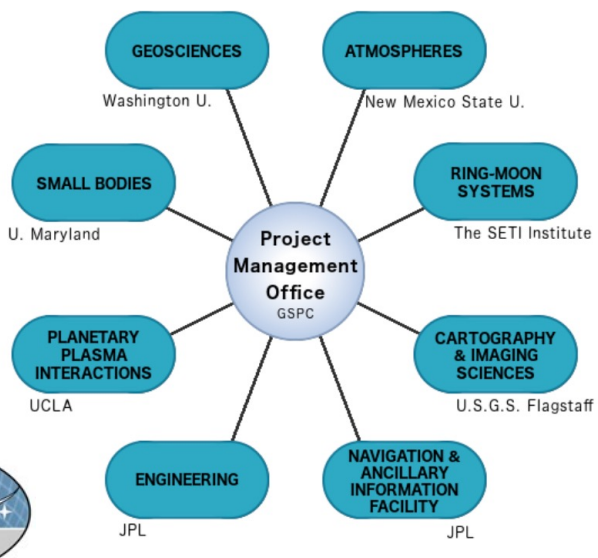


ads NASA

esa European Space Agency

NASA

IVOA



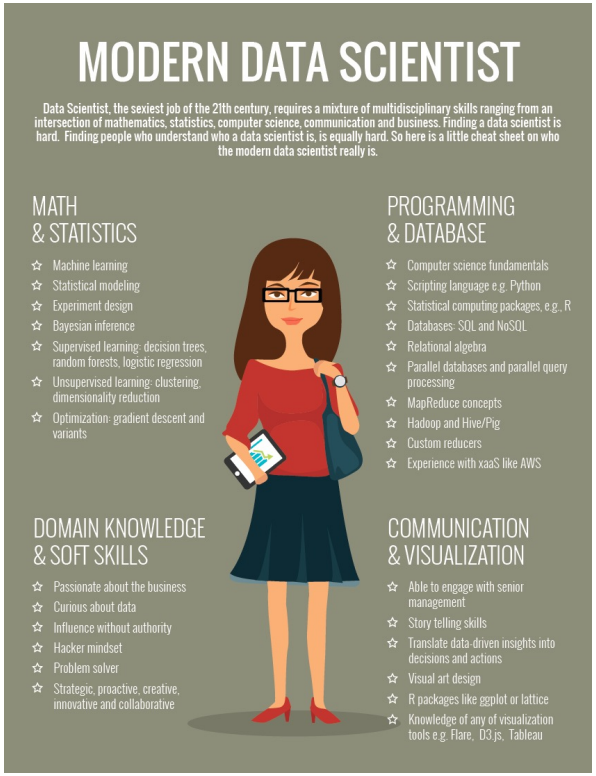
ALFRED P. SLOAN FOUNDATION 1934

NED NASA IPAC

EUROPEAN OPEN SCIENCE CLOUD

NSF

Infrastructure: Local and Social



https://github.com/okulbilisim/awesome-datascience

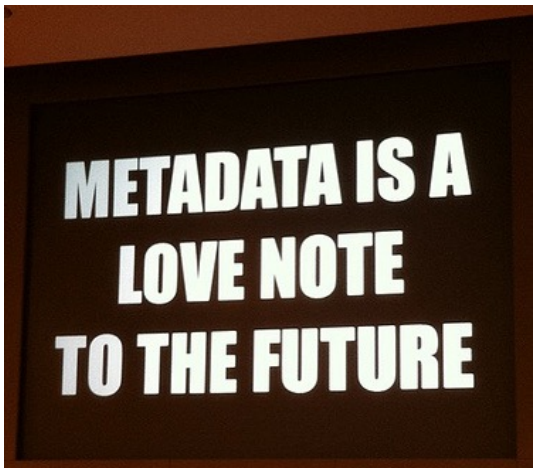
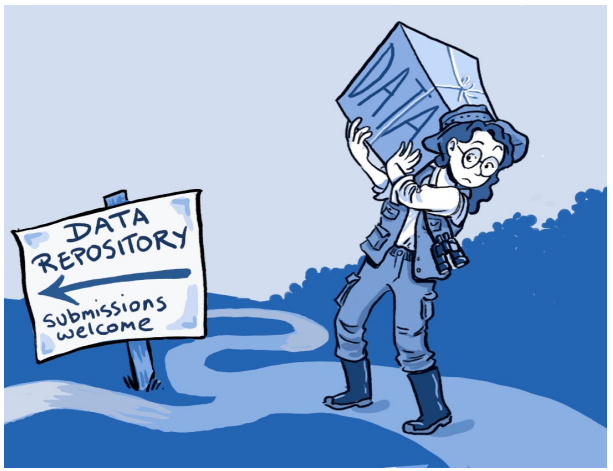


Photo by @kissane; presentation by Jason Scott (@textfiles)



https://en.wikipedia.org/wiki/Data_sharing

Photo Archive

The Getty Research Institute

CC Sean MacEntee, Flickr

Where are we now?

Image by Jean-Philippe Delberghe on Unsplash.com

Research Data Infrastructure components

- Technical
  - Networks, instruments
  - Computing capacity, software
- Institutional
  - ERICs
  - Data repositories
- Policy
  - Data management plans
  - Data sharing requirements
- Scholarly practices
  - Data sharing
  - Open access publishing

https://roadmap2021.esfri.eu

RESEARCH INFRASTRUCTURES TABLE

NAME	FULL NAME	TYPE	RIS TYPE
EBRAINS	European Brain ReseArch INfrastructureS	Distributed	project
SLICES	Scientific Large-scale Infrastructure for Computing/Communication Experimental Studies	Distributed	project
SoBigData++	European Integrated Infrastructure for Social Mining and Big Data Analytics	Distributed	project
PRACE	Partnership for Advanced Computing in Europe	Distributed	landmark
IFMIF-DONES	International Fusion Materials Irradiation facility - DEMO Oriented NEutron Source	Single-sited	project
MARINERG-i	Marine Renewable Energy Research Infrastructure	Distributed	project
ECCSEL ERIC	European Carbon Dioxide Capture and Storage Laboratory Infrastructure	Distributed	landmark
EU-SOLARIS	European Solar Research Infrastructure for Concentrated Solar Power	Distributed	landmark

IFMIF-DONES	International Fusion Materials Irradiation facility - DEMO Oriented NEutron Source	Single-sited	project
MARINERG-i	Marine Renewable Energy Research Infrastructure	Distributed	project
ECCSEL ERIC	European Carbon Dioxide Capture and Storage Laboratory Infrastructure	Distributed	landmark
EU-SOLARIS	European Solar Research Infrastructure for Concentrated Solar Power	Distributed	landmark
JHR	Jules Horowitz Reactor	Single-sited	landmark
DANUBIUS-RI	International Centre for Advanced Studies on River-Sea Systems	Distributed	project
DiSSCo	Distributed System of Scientific Collections	Distributed	project
eLTER RI	Integrated European Long-Term Ecosystem, critical zone and socio-ecological system Research Infrastructure	Distributed	project
ACTRIS	Aerosol, Clouds and Trace Gases Research Infrastructure	Distributed	landmark
EISCAT_3D	Next generation European Incoherent Scatter radar system	Single-sited	landmark
EMSO ERIC	European Multidisciplinary Seafloor and water-column Observatory	Distributed	landmark
EPOS ERIC	European Plate Observing System	Distributed	landmark
INSTRUCT ERIC	Integrated Structural Biology Infrastructure	Distributed	landmark
-RIHS	European Research Infrastructure for Heritage Science	Distributed	project
HRI	European Holocaust Research Infrastructure	Distributed	project
GP	The Generations and Gender Programme	Distributed	project
UIDE	Growing Up In Digital Europe: EuroCohort	Distributed	project
PERAS	OPen scholarly communication in the European Research Area for Social Sciences and Humanities	Distributed	project
ESILIENCE	REligious Studies Infrastructure: tooLs, Innovation, Experts, conNections and Centres in Europe	Distributed	project
ESSDA ERIC	Consortium of European Social Science Data Archives	Distributed	landmark
LARIN ERIC	Common Language Resources and Technology Infrastructure	Distributed	landmark
ARIAH ERIC	Digital Research Infrastructure for the Arts and Humanities	Distributed	landmark
SS ERIC	European Social Survey	Distributed	landmark

Methods to Share Data

- Deposit datasets in a data archive
- Publish data documentation
  - Research protocols
  - Codebooks
  - Software
  - Algorithms
- Link datasets to journal article or publication
- Cite data and software

National Institutes of Health Data Sharing Policy 2023

Section II. Definitions

For the purposes of the DMS Policy, terms are defined as follows:

SCIENTIFIC DATA	*The recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens.*
DATA MANAGEMENT	*The process of validating, organizing, protecting, maintaining, and processing scientific data to ensure the accessibility, reliability, and quality of the scientific data for its users.*
DATA SHARING	*The act of making scientific data available for use by others (e.g., the larger research community, institutions, the broader public), for example, via an established repository.*
METADATA	*Data that provide additional information intended to make scientific data interpretable and reusable (e.g., date, independent sample and variable construction and description, methodology, data provenance, data transformations, any intermediate or descriptive observational variables).*
DATA MANAGEMENT AND SHARING PLAN (PLAN)	*A plan describing the data management, preservation, and sharing of scientific data and accompanying metadata.*

Scientific Data:

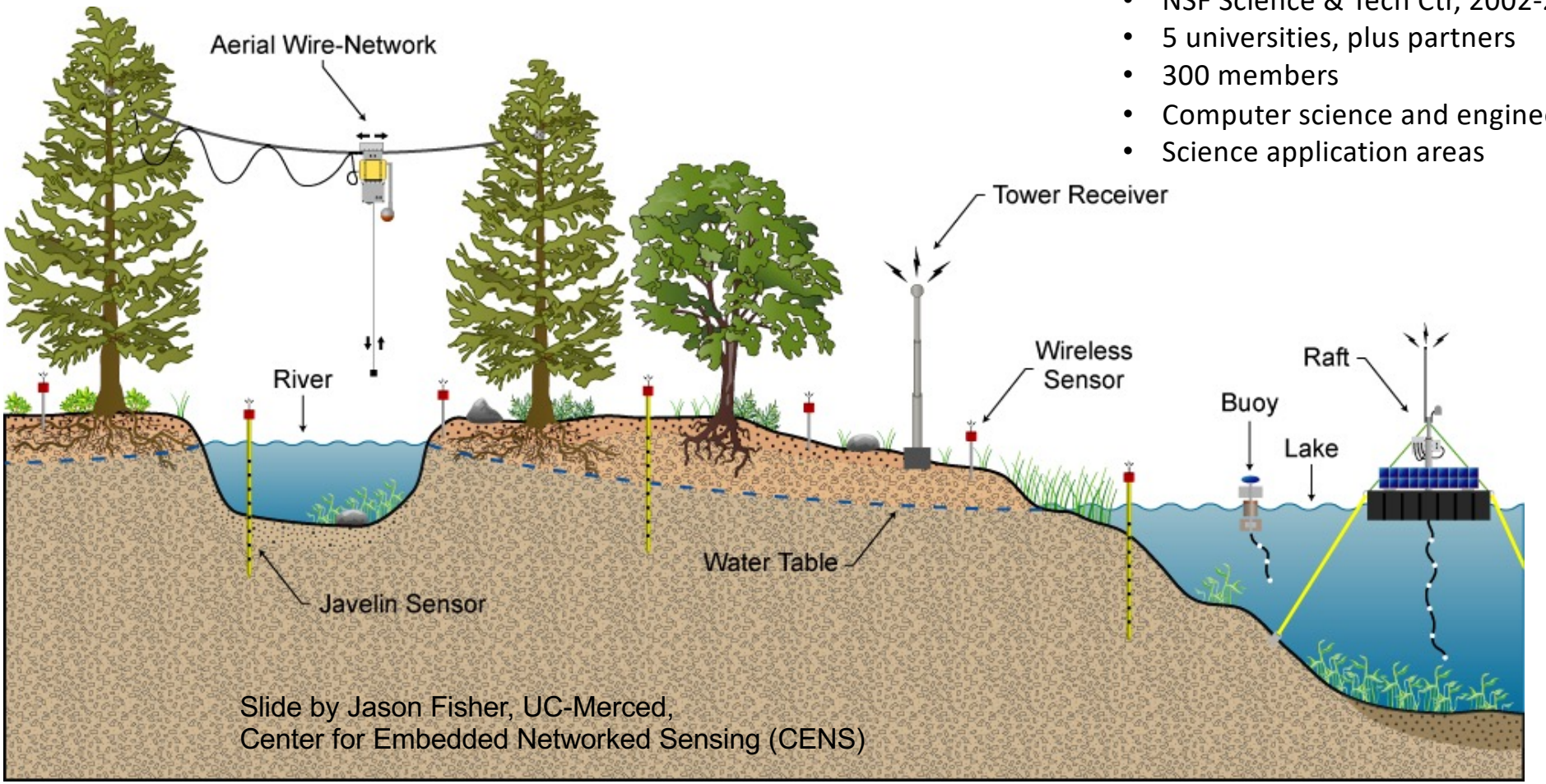
The recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens.

How did we get here?

Image by Jean-Philippe Delberghe on Unsplash.com

Center for Embedded Networked Sensing

- NSF Science & Tech Ctr, 2002-2012
- 5 universities, plus partners
- 300 members
- Computer science and engineering
- Science application areas

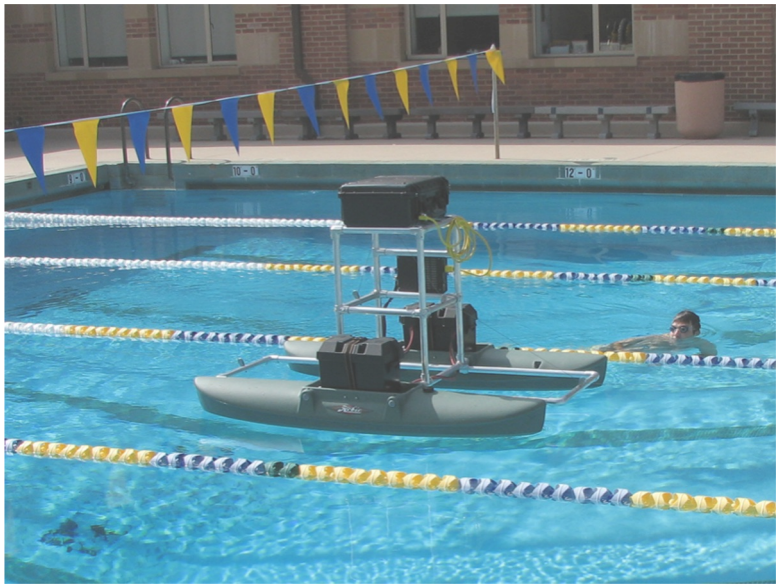


Slide by Jason Fisher, UC-Merced,
Center for Embedded Networked Sensing (CENS)

Science <–> Data

Engineering researcher:

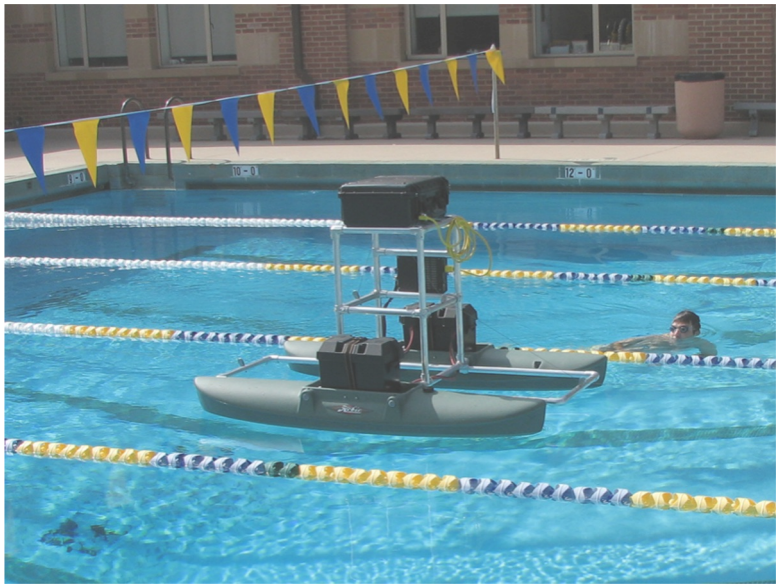
"Temperature is temperature."



CENS Robotics team

Science <–> Data

Engineering researcher:
"Temperature is temperature."



CENS Robotics team

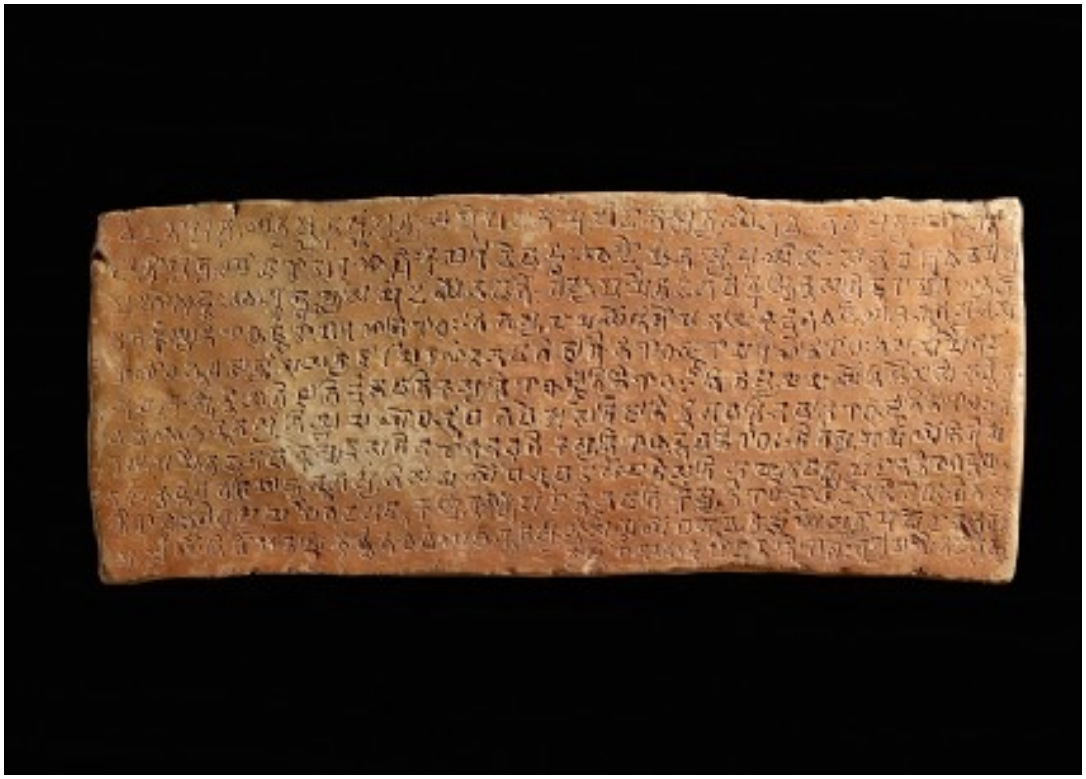
Biologist*: **"There are hundreds of ways to measure temperature.***
'The temperature is 98' is low-value compared to, 'the temperature of the surface, measured by the infrared thermopile, model number XYZ, is 98.' That means it is measuring a proxy for a temperature, rather than being in contact with a probe, and it is measuring from a distance. The accuracy is plus or minus .05 of a degree. I [also] want to know that it was taken outside versus inside a controlled environment, how long it had been in place, and the last time it was calibrated, which might tell me whether it has drifted.."

Opening a box of data: Chinese Buddhist Philology

Stefano Zacchetti
Yehan Numata Professor
of Buddhist Studies
Oriental Institute
University of Oxford

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. MIT Press.

Bricks in the wall…

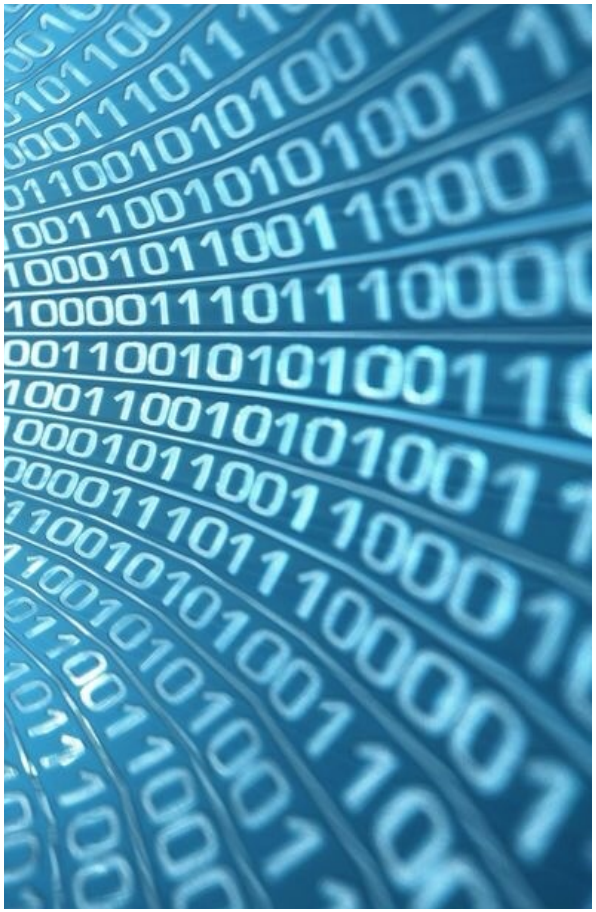


Brick inscribed with the Sutra on Dependent Origination *Gorakhpur district, late 5th century - early 6th century AD. Ashmolean Museum*

What challenges do we face now?

Image by Jean-Philippe Delberghe on Unsplash.com

Challenges for Research Data Infrastructure



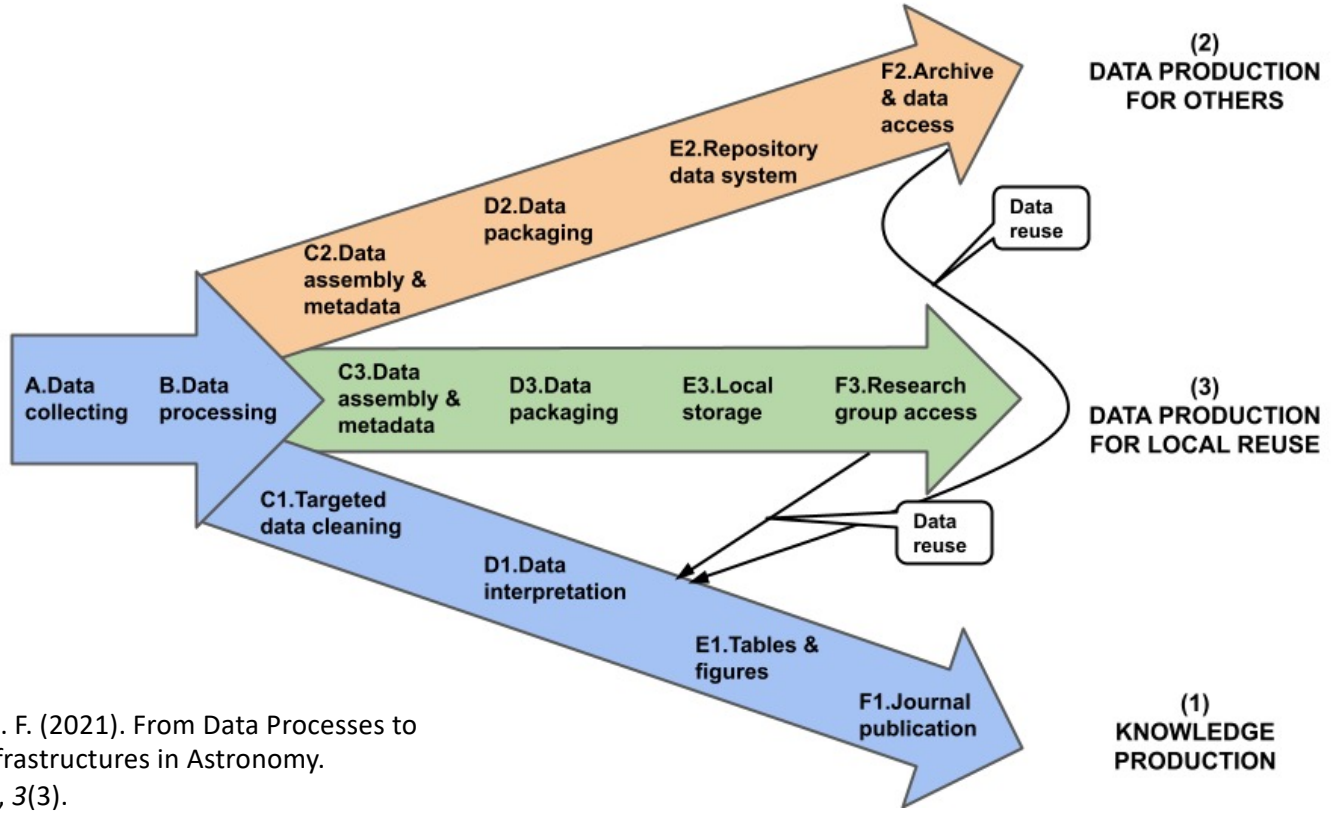
- How to
  - decide what data are worth keeping?
  - make data useful and reusable?
  - balance costs and benefits?
  - balance incentives and risks?
  - steward data resources?
  - govern research data resources?
  - pay for infrastructure?

Lack of incentives to share data

- Labor to document data
- Benefits to unknown others
- Competition
- Control
- Confidentiality
- Lack of expertise and staff
- Lack of sustainability…

Image source: www.buildingsrus.co.uk/…/ target1.htm

Data production, knowledge production, and reuse



- Borgman, C. L., & Wofford, M. F. (2021). From Data Processes to Data Products: Knowledge Infrastructures in Astronomy. *Harvard Data Science Review*, *3*(3).
- Baker, K. S., & Mayernik, M. S. (2020). Disentangling knowledge production and data production. *Ecosphere*, *11*(7), e03191.

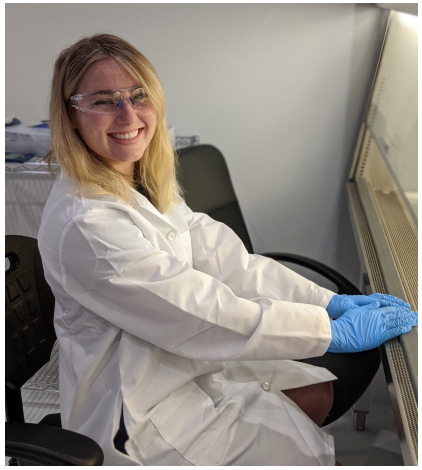
Data Creators' Advantage

Comparative Data Reuse

- Ground truthing: calibrate, compare, confirm
- Instrument calibration
- Frequent, routine practice

Integrative Data Reuse

- Analysis: identify patterns, correlations, causal relationships
- Novel statistical analyses
- Rare, emergent practice



Bret Kavanaugh, Unsplash

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and reuses of scientific data: The data creators' advantage. *Harvard Data Science Review,* 1:2

National Cancer Institute

Software Creators' Advantage

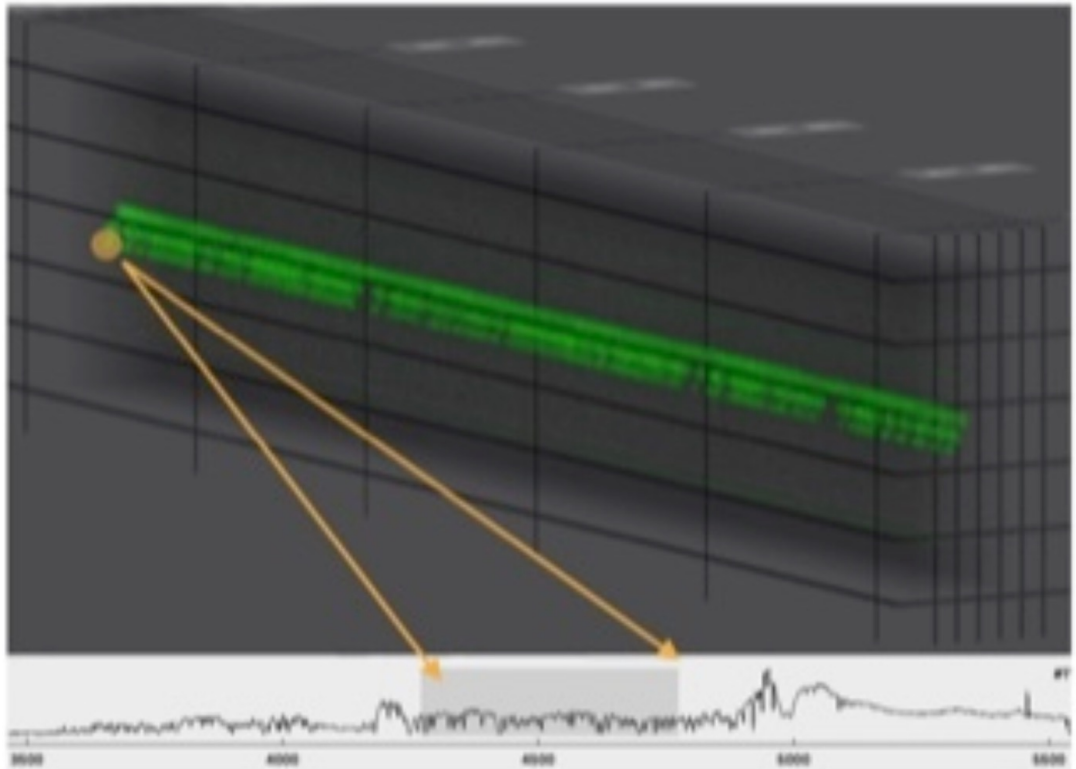
Software Heritage

Software is fragile

unlike words carved in stone it can be deleted or get corrupted

W.M. Keck Telescopes, Data Reduction pipeline for the Cosmic Web Imager, Infrared Processing and Analysis Center, Caltech

The installation and usage of the DRP is described in https://kcwi-drp.readthedocs.io/en/latest. The DRP delivers science quality products and includes geometric, wavelength, and flux calibration. It can run completely unattended (including during the observing run at Keck Observatory) but it also offers a number of options to customize the reduction according to the specific science needs.



A portion of a KCWI data cube of a gravitationally lensed quasar pair. The data cube was produced by the new pipeline. The lower inset shows a spectrum drawn from an aperture placed over a portion of the cube encompassing one of the quasar images.

Data creation and reuse: The Ideal

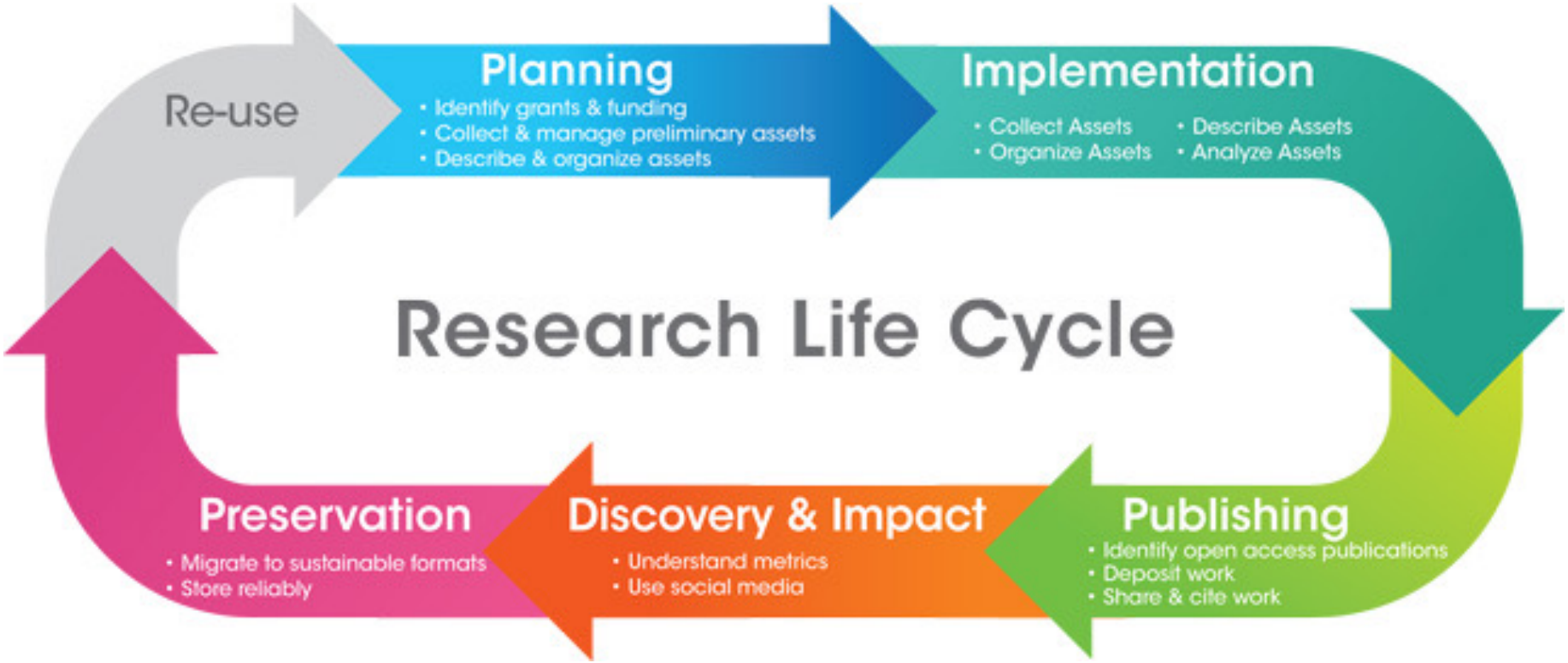
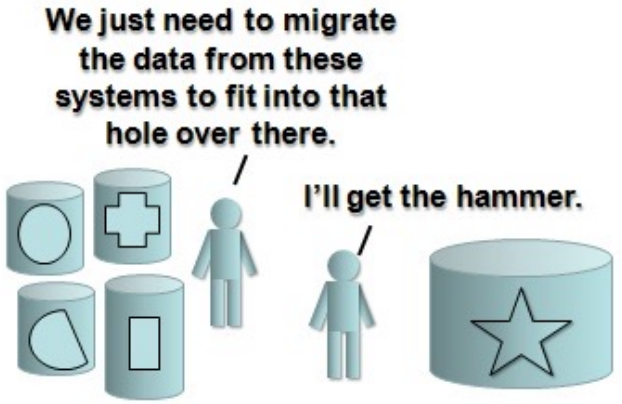


Image: http://www.lib.uci.edu/dss/images/lifecycle.jpg

Borgman, C. L. (2019). The lives and after lives of data. *Harvard Data Science Review*, *1*(1).

Data Stewardship: The Reality

http://www.information-age.com/cloud-computing-pharmaceutical-industry-123462676/



http://www.datamartist.com/data-migration-part-1-introduction-to-the-data-migration-delema

Mount Wilson Solar Observatory, 2017

Getty Research Institute

http://gsa.rice.edu/

Graduate students

https://med.nyu.edu/our-community/life-nyu-school-medicine/life-postdoc

Post-doctoral fellows

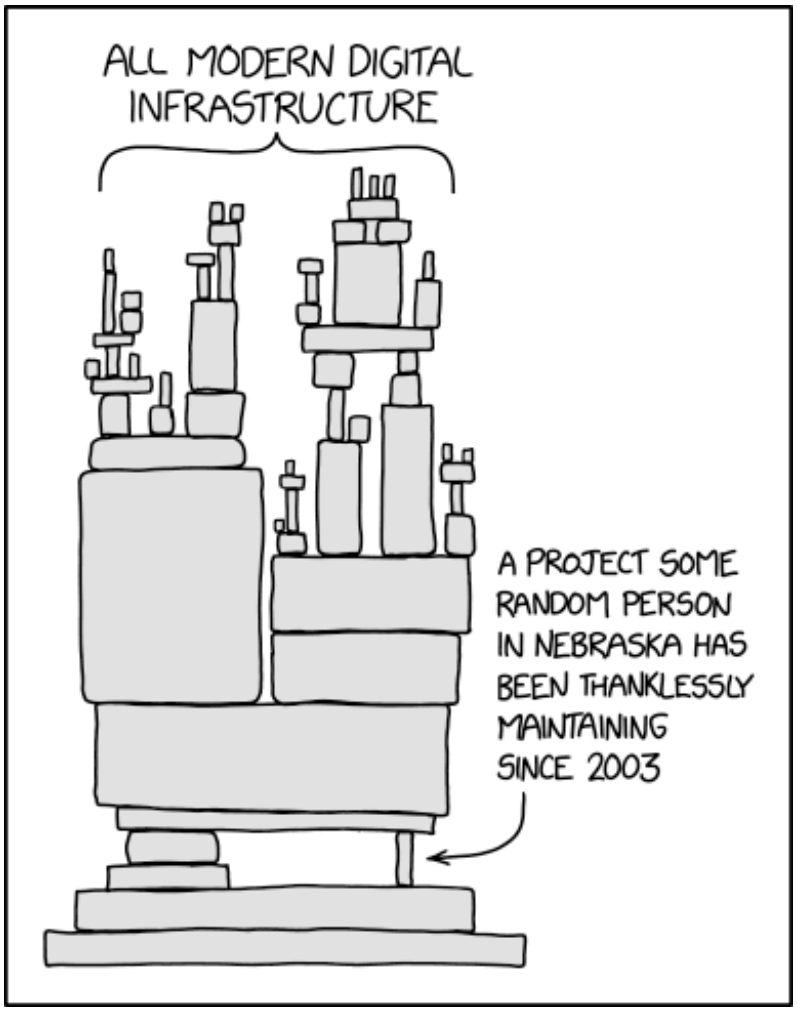
Where to go from here?

Image by Jean-Philippe Delberghe on Unsplash.com

Infrastructure: Fragility

- Brittleness
  - Maintenance and repair
  - Invisible until breakdown
  - Changes in installed base
- Human resources
  - Data stewardship
  - Skill sets, Help desks
  - Local and global communities
- Interoperability
  - Hardware, software, networks
  - Language
  - Instrumentation
- Risks
  - Cyberattacks
  - Misuse, appropriation
  - Confidential, proprietary information

Borgman, Darch, Sands, & Golshan (2016). The durability and fragility of knowledge infrastructures. *ASIST Proc*, *53*, 1–10.



https://www.explainxkcd.com/wiki/index.php/2347:_Dependency

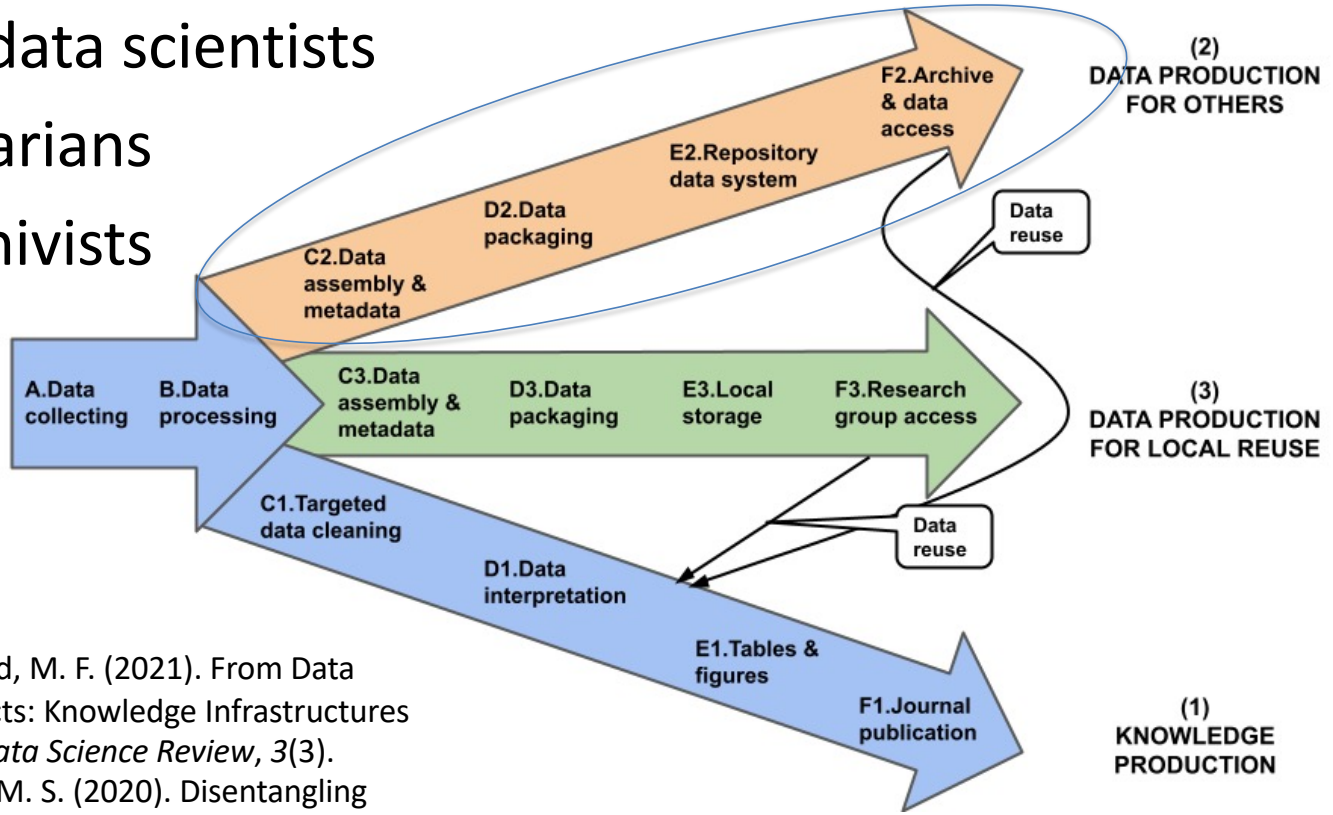
Infrastructure: Durability

- Collaboration and openness
- International coordination
- Long-term value of data
- Agreed standards
  - Units of measurement
  - Data structures
- Shared resources
  - Missions, instruments
  - Data archives
  - Tools and technologies
- Maintenance commitments

Borgman, Darch, Sands, & Golshan (2016). The durability and fragility of knowledge infrastructures. *ASIST Proc*, *53*, 1–10.

Data Management Workforce

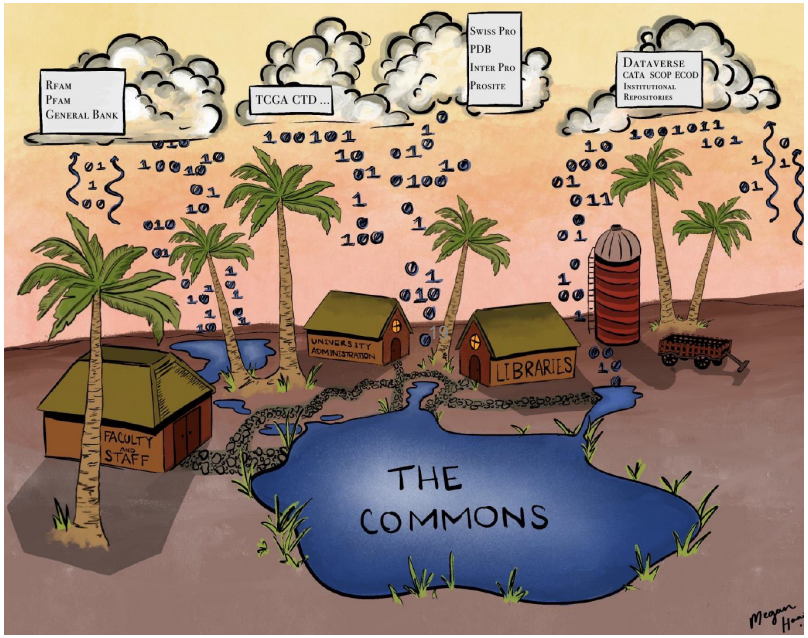
- Domain data scientists
- Data librarians
- Data archivists



- Borgman, C. L., & Wofford, M. F. (2021). From Data Processes to Data Products: Knowledge Infrastructures in Astronomy. *Harvard Data Science Review*, *3*(3).
- Baker, K. S., & Mayernik, M. S. (2020). Disentangling knowledge production and data production. *Ecosphere*, *11*(7), e03191.

Governance: Building the Village

- Data sharing is a 'collective action problem'
- Holistic approaches to sharing infrastructure
  - Distribute responsibility among stakeholders
  - Invest in data management expertise
  - Reframe goals in collective terms
- Fund the commons
  - Public support for data repositories
  - International exchange of best practices
- Invest in sustainable strategies



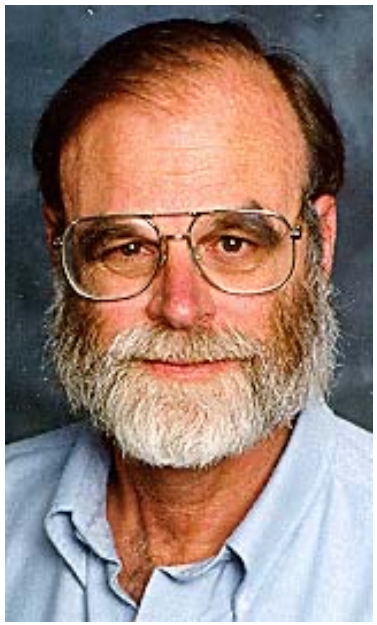
Borgman, C. L., & Bourne, P. E. (2022). Why it takes a village to manage and share data. *Harvard Data Science Review.* Illustration by Megan Haas

Data, Infrastructure, and Stewardship

- Whose data?
  - Global, comparative, fungible
  - Local, integrative, specific
- Whose infrastructure?
  - Funders, universities, companies
  - Individual investigators
- Whose stewardship?
  - Maintain collections, models, instruments, technology, code…
  - Invest in people, skills, collaborations

May all your problems be technical

Jim Gray, Turing Award Winner



Acknowledgements: Talk Preparation

- Amy Brand, MIT Press
- Alyssa A. Goodman, Harvard University, Astronomy
- Peter T. Darch, U of Illinois, Information Sciences
- Matthew S. Mayernik, National Center for Atmospheric Research
- Irene V. Pasquetto, U of Maryland, Information Studies

References

- Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. MIT Press.
- Borgman, C. L. (2019). The lives and after lives of data. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.9a36bdb6
- Borgman, C. L., & Bourne, P. E. (2022). Why It Takes a Village to Manage and Share Data. *Harvard Data Science Review*, *4*(3). https://doi.org/10.1162/99608f92.42eec111
- Borgman, C. L., & Brand, A. (2022). Data blind: Universities lag in capturing and exploiting data. *Science*, *378*(6626), 1278–1281. https://doi.org/10.1126/science.add2734
- Borgman, C. L., Darch, P. T., Sands, A. E., & Golshan, M. S. (2016). The durability and fragility of knowledge infrastructures: Lessons learned from astronomy. *Proceedings of the Association for Information Science and Technology*, *53*, 1–10. http://dx.doi.org/10.1002/pra2.2016.14505301057
- Borgman, C. L., & Wofford, M. F. (2021). From Data Processes to Data Products: Knowledge Infrastructures in Astronomy. *Harvard Data Science Review*, *3*(3). https://doi.org/10.1162/99608f92.4e792052
- Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. MIT Press.
- Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and Reuses of Scientific Data: The Data Creators' Advantage. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.fc14bf2d