Data Practices

Data Practices

An empirical view of what people creating, analyzing, and managing data *actually do*.
(or would do)
so that we can improve efficiency and reliability

V1. Data Practices	(how do we know what works?)
V2: What’s going on in the lab?	(brace yourself; it ain’t pretty)
V3: Data sharing	(no, no, no, no, no. It’s *mine*!)
V4: Data Reuse	(if you didn’t make it, it is hard to use it)

V3: Data sharing

Why data sharing is important

Why data sharing is hard

Impediments to sharing

Incentives?

Why is data sharing important

Good data is, course, important and *valuable to communities beyond the developing community*

And it is arduous, time-consuming, and expensive to develop

And often we need relevant data immediately (crisis informatics)

So failure to share creates lost opportunity and additional expense

And can have extremely serious consequences

(consider data in medicine, engineering, etc.
or data needed to address a disaster, such as a hurricane)

Why data sharing is hard

On **the receiving side** there are of course the usual *data integration* difficulties:

finding relevant data,
getting needed permissions and licenses,
and integrating data in different formats and description standards into the receiving system
of applications tools and practices – and it much be correctly understood.

But we've already had a good look at these problems earlier, and we'll be revisiting some of them again in the next video, on data Reuse.

Here we focus on **the sharing side**, that is: why share?

Data misuse concerns

Question	Agree
Data may be misinterpreted due to complexity of the data.	75% (n=1293)
Data may be used in other ways than intended.	74% (n=1289)
Data may be misinterpreted due to poor quality of the data.	71% (n=1291)

(Tenopir et al., 2011)

More data sharing impediments

Astronomy (Gray et al)

- Laborious process
- Few standards

Science & Humanities (Borgman)

- Laborious effort
- No rewards for sharing data
- Lose competitive advantage
- Data ownership

Other challenges

- Grant cycles & funding
- Domains without repositories
- Concerns of data misuse
- Legal and ethical issues
- Co-authorship expectations
- (dis)incentives – tenure, promotion

(Cragin et al., 2010; Tenopir et al., 2011)

What we hear is not encouraging

Where’s the best place for my data?

My data’s available/archived on…[my computer, server, website].

Of course I’m willing to share my data, but...

My data will never be of use to anyone else.

There are no standards in my field.

What version of the data should I share?
Raw vs. Processed, Continuously streaming data.

Researchers will need my special analysis tools to reuse the data.

Sharing practices vary by discipline

	Culture of data sharing	Infrastructure for data sharing	Effect of open data policies	Overall propensity to share data
Astronomy	High	Low	Medium	High
Chem. Crystallography	Medium	Low	Low	High
Genomics	High	Medium	High	High
Systems biology	Medium	High	High	Medium
Classics	High		Medium	Medium
Social/Public Health	Low	Low	Low	Low
RELU	Medium	Low	Medium	Medium
Climate science	Low	Low	Medium	Low to medium

Research Information Network (RIN), 2008

Incentives for sharing …?

Scientific value
- Better analysis and research outcomes

Credit
- Credit for data producers (metadata)
- Data sharing = increased citations (Pinowar, 2007)

Infrastructure [*this time from feedback*]
- Interoperable applications, systems, and data
- Reliability and reproducibility
- Efficiency
- Easier collaboration

Reciprocity
- You give some, you get some

Tenure and promotion assessment
- Measure of being a good data steward

References (General)

Ball, A. (2010). Data lifecycles. In Review of the State of the Art of the Digital Curation of Research Data. Project Report. Bath, UK: University of Bath.

Borgman, C.L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science & Technology,* 63(6): 1059-1078.

Babeu, A. (2011). "Rome Wasn't Digitized in a Day": Building a Cyberinfrastructure for Digital Classics. Washington DC.

Chao, T. C., Cragin, M. H., & Palmer, C. L. (2014). Data Practices and Curation Vocabulary (DPCVocab): An empirically derived framework of scientific data practices and curatorial processes. JASIST.

Hanson, K., Surkis, A., & Yacobucci, K. (2012). Data Sharing and Management Snafu in 3 Short Acts [video].

Hey, A. J., Tansley, S., & Tolle, K. M. (2009). The fourth paradigm: data-intensive scientific discovery.

Cragin, M.H., Palmer, C.L., Carlson, J.R., & Witt, M. (2010). Data sharing, small science, and institutional repositories. *Philosophical Transactions of the Royal Society A*, *368*(1926), 4023-4038.

Pepe, A., Goodman, A., Muench, A., Crosas, M. & Erdmann, C. (2014). How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. *PLoS ONE,* 9(8): e104798.

Research Information Network. (2008). To Share or Not to Share: Publication and Quality Assurance of Research Data Outputs.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., et al. (2011). Data sharing by scientists: Practices and perceptions. PloS ONE, 6(6), e21101.

Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, et al. (2015) Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide.