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)

V4: Data reuse

What is reuse and why is it important?

reuse vs sharing

What are the obstacles to reuse?

How can reuse be supported?

Data reuse, what is it

“[data reuse is] …the use of data collected for one purpose to study a new problem.”
(Zimmerman, 2008)

And perhaps also:

the use of data collected by one technical community
and being used by another technical community”

And, most often, both of those apply:

Different *communities* **and** different *problems*

[and so: different methods, practices, vocabularies, software . . . etc.]

(Tenopir et al., 2014)

Reuse vs Sharing

Data sharing and *data reuse* are closely related,
but a rough distinction can be drawn,
focusing on the perspective taken by research studies in this area.

Data sharing studies tend to focus on:

Why and how data producers share (or don’t share) their data
and how we can encourage and support data sharing

Data reuse studies tend to focus on:

How communities use (or why they don’t use) relevant data that they did not produce
and for which they are not the intended consumer*
and how we can encourage and support reliable and efficient reuse of such data

So data sharing focuses on the *producer*, and data reuse on the *consumer*.
Not surprisingly the respective issues tend to be mirror images.

*cf. “designated community” (OAIS).

Why data ~~sharing~~ reuse is 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~~ ^reuse 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)

Some empirical data on consequences

	Mean, 5-1 Scale Mean (Std. Dev.)*
Lack of access to data generated by other researchers is a major impediment to progress in science.	3.99 (1.03)
Lack of access to data generated by other researchers has restricted my ability to answer scientific questions.	3.36 (1.27)

* 1 = Strongly disagree to 5 = Strongly agree

(Tenopir et al., 2014)

Reuse challenges

Factors influencing reuse of a data set:

Discovering the data

Assessing relevance

Serialization and file format issues

logical and conceptual modeling issues

Semantic variations (vocabularies, definitions of terms, etc)

Trustworthiness (inputs, algorithms, provenance, workflow)

Intellectual property, credit, and regulatory issues

and more.

Obviously many of these are similar to data integration issues discussed earlier

And, equally obvious:

On the supply side **standards** and **documentation** (especially metadata) are key to supporting reuse.

Area-specific challenges to data reuse (and sharing)

Medical, financial, social, governmental

Heavily regulated by federal and state statutes and common law.

Major tort and statutory liabilities for violation.

Security requirements for allowed use may be unavailable and can be expensive.

and so on

Other for-profit industry (in addition to challenges above)

Data or data access may be revenue-generating business product.

Data has strategic value (or vulnerability) to unit
with negative consequences (even if greater social value) if public.

Even if data is available provenance and workflow (etc) information may be restricted,
limiting value of data

Data circulation can trigger restraint of trade (e.g. price fixing, Sherman Antitrust Act).

Licensing violations can create substantial financial vulnerability.

and so on

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.