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)

V1. Data Practices

How do we know what works?

[The *empirical science* of data curation]

Empirical studies

Constructed, Naturalistic.

Data collection, data analysis

The empirical science of data curation

The science of *data analytics* is interdisciplinary
(and partly social science)

but mostly **mathematics**

The science of *data curation* is interdisciplinary
(and partly mathematics)

but mostly **social science**

A fundamental question in the science of data curation is

How can we more efficiently and reliably support the use of data?

This is clearly an *empirical* question

The empirical science of data curation: How to we do it?

How can we more efficiently and reliably support the use of data?

To answer this question we must conduct *empirical* research.

We can divide empirical research in this area into two rough categories

Constructed studies

Naturalistic studies

Constructed studies

These typically follow the classic scientific model. . .

conjecturing a hypothesis
and constructing an experiment or other targeted data collection to elicit confirmation.

Constructed studies are especially useful for
resolving an issue about what is influencing outcomes,
determining how a particular intervention might affect outcomes,
testing a new tool or practice.

Naturalistic studies

These typically collect data about an actual ongoing research situation,

observing what researchers and their staff actually are doing,
and then using that data to develop theories about data curation,
or ideas for changes in practices.
[There may or may not be an hypothesis at the outset.]

Naturalistic studies can be useful for identifying problems and opportunities
or developing general picture of a research and data management practices in a field.

Methods

Data collection methods in empirical studies of research and data curation processes include:

interviews, surveys, transaction log analysis, work product analysis (code, data, workflow) time and motion studies, experiments, simulations, and so on.

Analytical methods in empirical studies of research and data curation processes include

mathematics, qualitative analysis, interpretative methods, often informed by theories and results in computer science, information science, and cognitive science.

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.

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)

V2: What's going on in the lab?

Empirical extraction of vocabulary and processes

Empirical identification of bad behavior:

- metadata?? (for retrieval? use? interpretation? preservation? credit? reproducibility?)
- code documentation?
- code testing?
- workflow documentation?
- provenance availability?

and so on

Incentives to do better?

The problem (we're human, all too human)

Quasi-empirical studies

Much of the global analysis of research processes, data lifecycles, and data curation is

basically empirical,

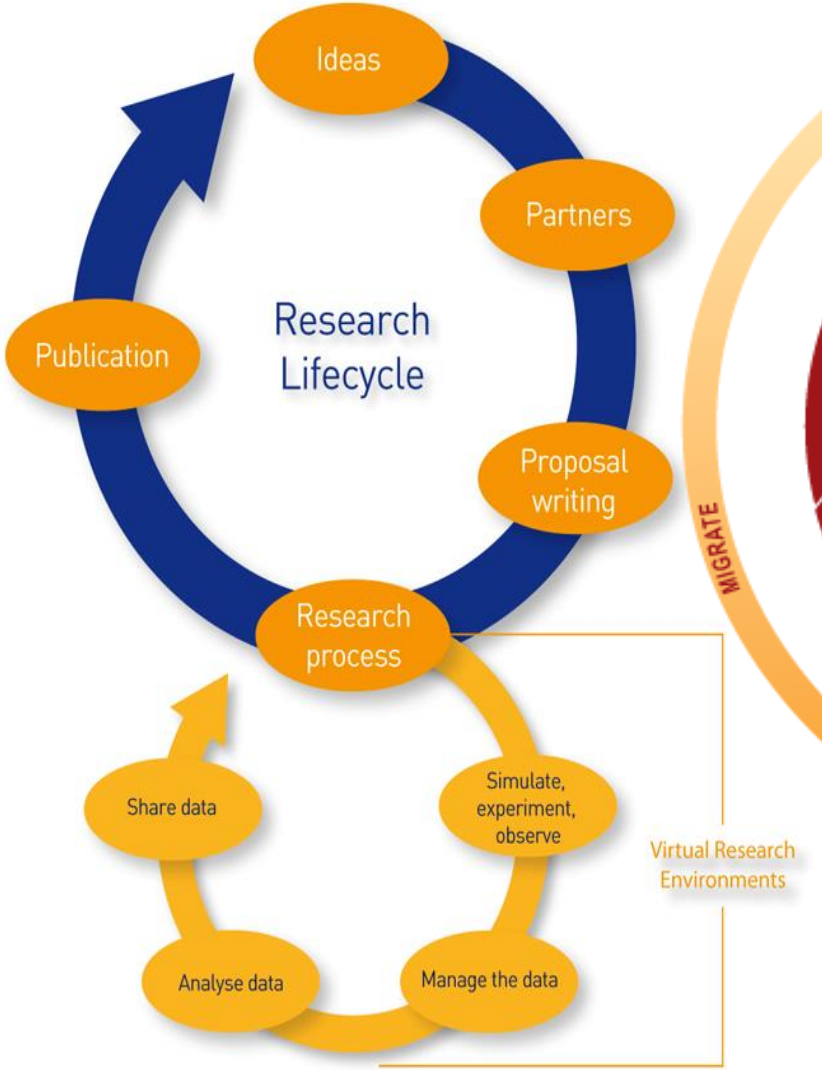
but at the same time casual, not rigorous

That does mean it is wrong;

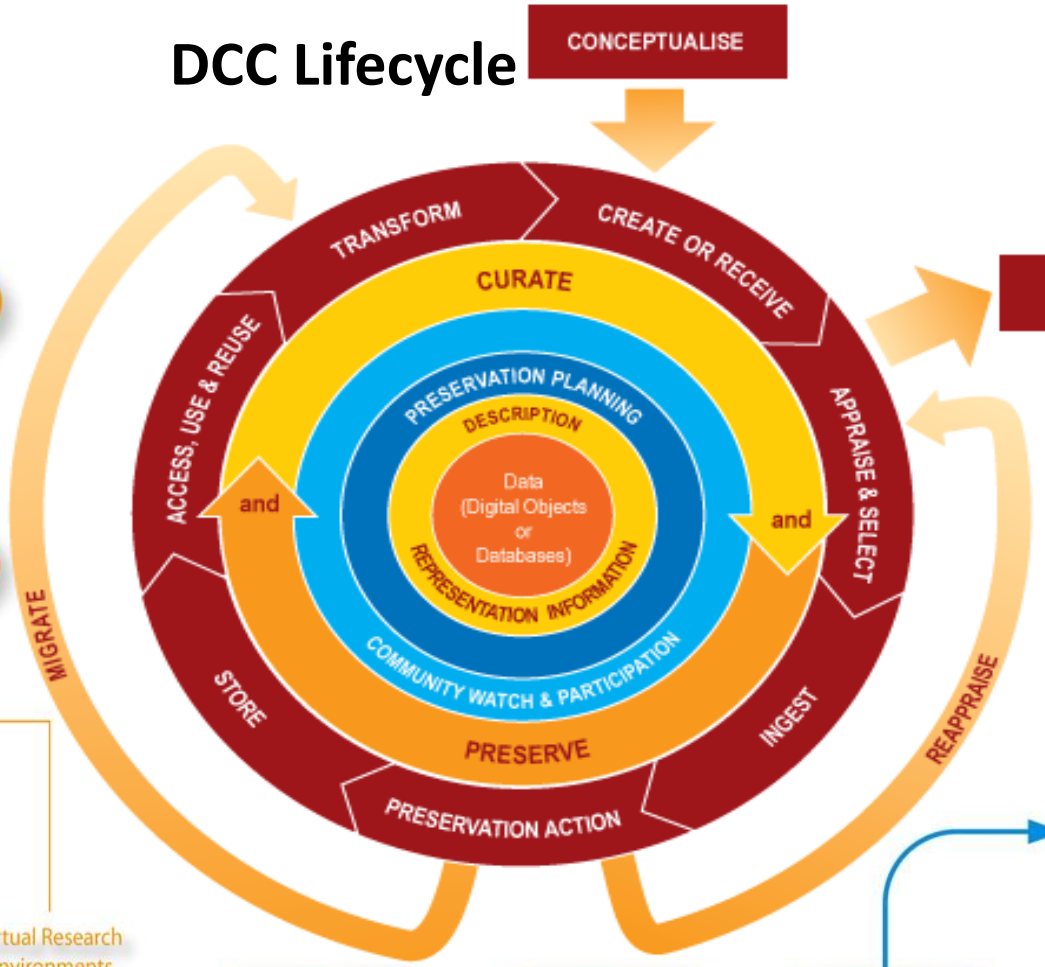
on the contrary: we don’t always need a rigorous designed study

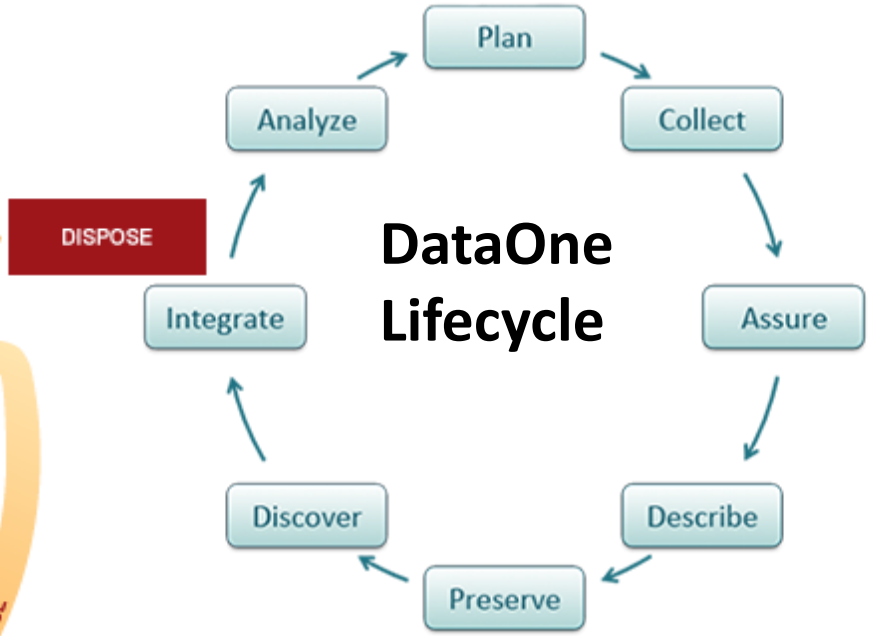
For instance:

JISC/VRE Lifecycle



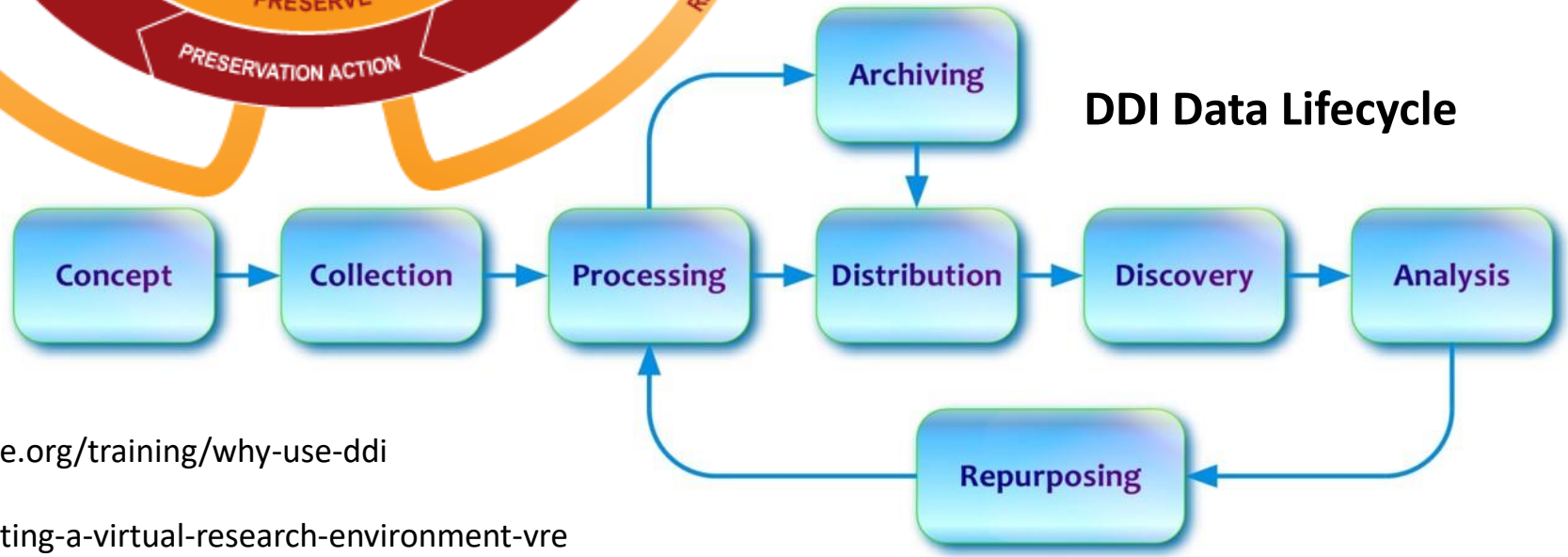
DCC Lifecycle





DataOne Lifecycle

DDI Data Lifecycle



Data Documentation Initiative. http://www.ddialliance.org/training/why-use-ddi
DataONE. https://www.dataone.org/data-life-cycle
JISC/VRE: https://www.jisc.ac.uk/full-guide/implementing-a-virtual-research-environment-vre

An empirically derived typology of research data practices

Designing research

Managing data

Generating and collecting

Processing

Analyzing, interpreting, and abstracting

Representing data

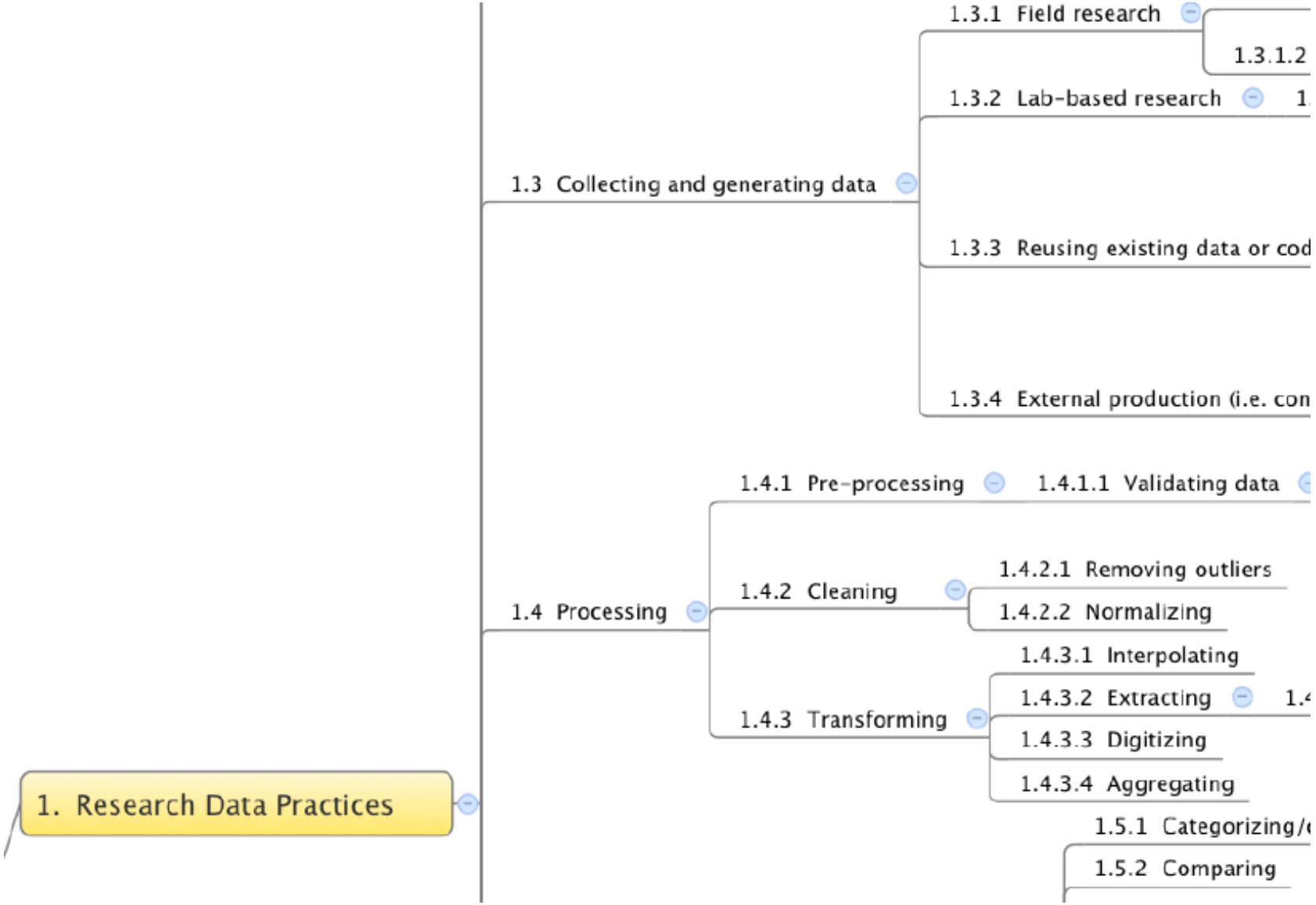
Sharing data and products

Attributing and citing data

Publishing data

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.

A fragment of an empirically derived “data practices and curation vocabulary” (DPCVocab)

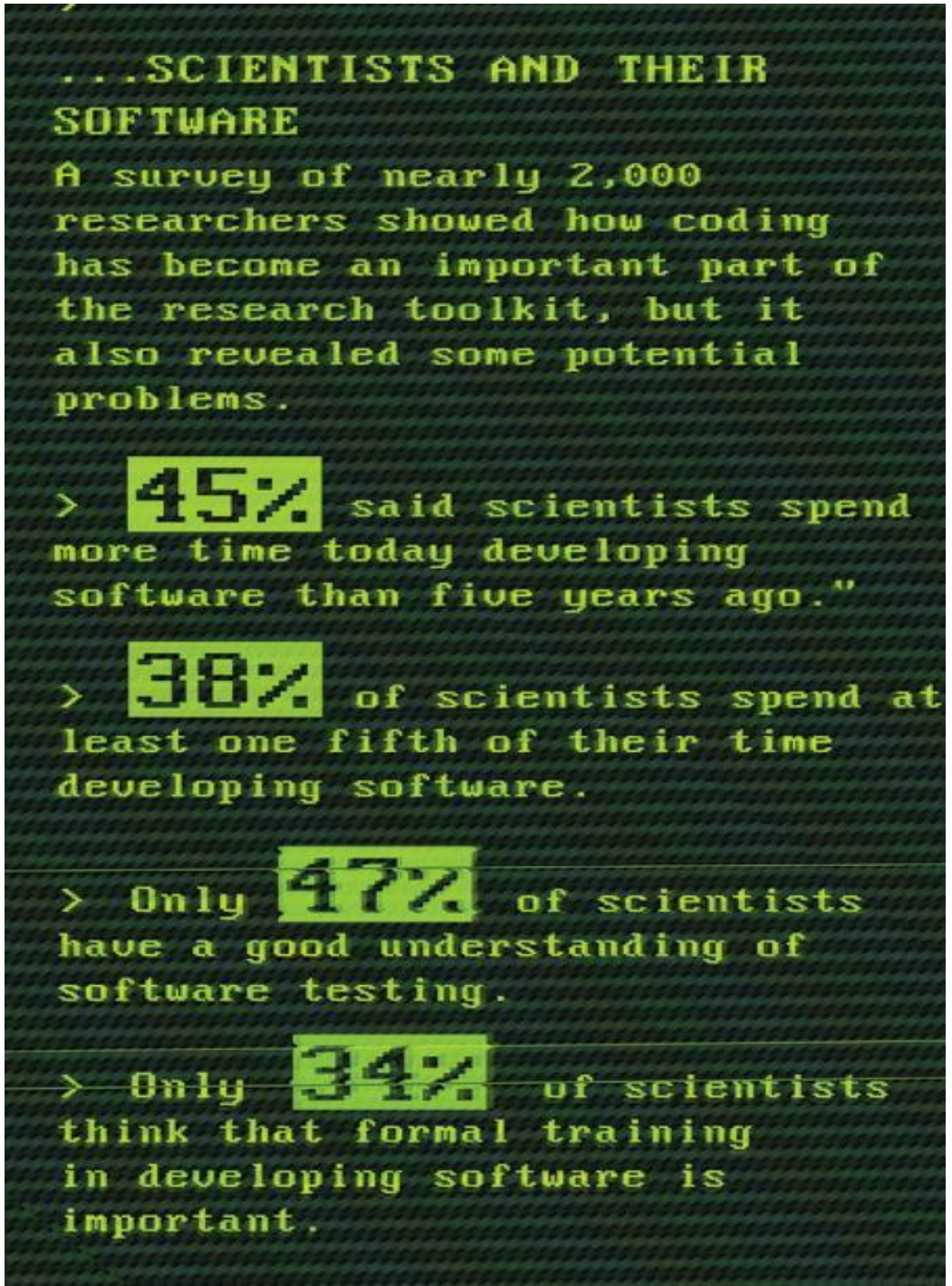


Chao, T. C., Cragin, M. H., & Palmer, C. L. (2014). Data Practices and Curation Vocabulary (DPCVocab): An empirically derived framework of scientific data practices and curatorial processes. JASIST. DOI: 10.1002/asi.

Discovering (ok, confirming) bad behavior

"As a general rule, researchers do not test or document their programs rigorously, and they rarely release their code, making it almost impossible to reproduce and verify published results generated by scientific software, say computer scientists. ... scientists often lack these communication and documentation skills”

— Zeya Merali “Computational science: ...Error . Why scientific programming does not compute”
Nature news feature (2010);
http://www.nature.com/news/2010/101013/full/467775a.html



Metadata failures

What metadata do you currently use to describe your data, if any?

Standards	2014 Responses
DC (Dublin Core)	7.1%
DwC (Darwin Core)	2.0%
DIF (Directory Interchange Format)	1.7%
EML (Ecological Metadata Language)	9.3%
FGDC (Federal Geographic Data Committee)	8.5%
ISO 19115 (Geographic Information-Metadata)	10.2%
OGIS (Open GIS)	7.2%
Standard within my lab	16.7%
Other	8.6%

None: 47.9%

(Tenopir et al., 2014)

Data storage

How much of your data do you currently store in the following locations?

	Most or all of my data
External hard disk/drive storage	83.3%
On my personal computer	65.3%
Dropbox/Google/Figshare/Cloud	57.2%
On my institution's server	37.7%
On the PI's server	28.4%
On a departmental server	23.1%
On paper in my office	13.7%
In my institution's repository	11.3%
In a domain repository	9.5%
Other data repository or archive	9.3%
In a publisher repository	2.4%

(Tenopir et ala., 2014)

Why is it so hard to be good?

We don't need a behavioral economist
to tell us that we have have a hard time giving up short-term benefits for long-term benefits,
even when the long-term benefits are greater.

And that's when the benefits accrue to *ourselves* (our future selves).

How much harder it is when much of the benefit accrues to *others*

This is why we don't document code, test our code, add metadata to datasets, use standards, backup our files, avoid transformations at the command line, etc. etc.

Often the benefits seems indirect and elusive, and we can convince ourselves it is unnecessary

["No time to document this, but no need either: how it works is self-evident.
And no need to test it: we were careful.
And no need to back up an earlier version; this one is better,
and I don't think we used that earlier version for anything important . . . (or did we?)"]

Incentives for good data practices …?

Scientific value
- Better analysis and research outcomes

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

Infrastructure
- Interoperable applications, systems, and data
- Reliability and reproducibility
- Efficiency
- Easier collaboration

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.

References (Lifecycle models)

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.

DataONE. Data Lifecycle Model. https://www.dataone.org/data-life-cycle

Digital Curation Centre. DCC Curation Lifecycle Model. http://www.dcc.ac.uk/resources/curation-lifecycle-model

ICPSR. (2012). Guide to Social Science Data Preparation and Archiving. 5th Edition. https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/

UCF Libraries. Resarch Lifecycle. http://library.ucf.edu/about/departments/scholarly-communication/research-lifecycle/

UK Data Archive. Create and Manage Data – Research Data Lifecycle. http://data-archive.ac.uk/create-manage/life-cycle

USGS. The Data Lifecycle. https://www2.usgs.gov/datamanagement/why-dm/lifecycleoverview.php

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.

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.