Mercè Crosas
Director of Data Science, IQSS
Harvard University
Twitter: mercecrosas

10 SIMPLE RULES FOR THE CARE AND FEEDING OF SCIENTIFIC DATA

Based on:
Goodman, Blocker, Borgman, Cranmer, Crosas, Di Stefano, Gil, Groth, Hogg, Kashyap, Hedstrom, Mahabal, Siemiginowska, Slavkovic, Pepe; 2013, PLOS, *In Review*

Projects by the Data Science Team at IQSS

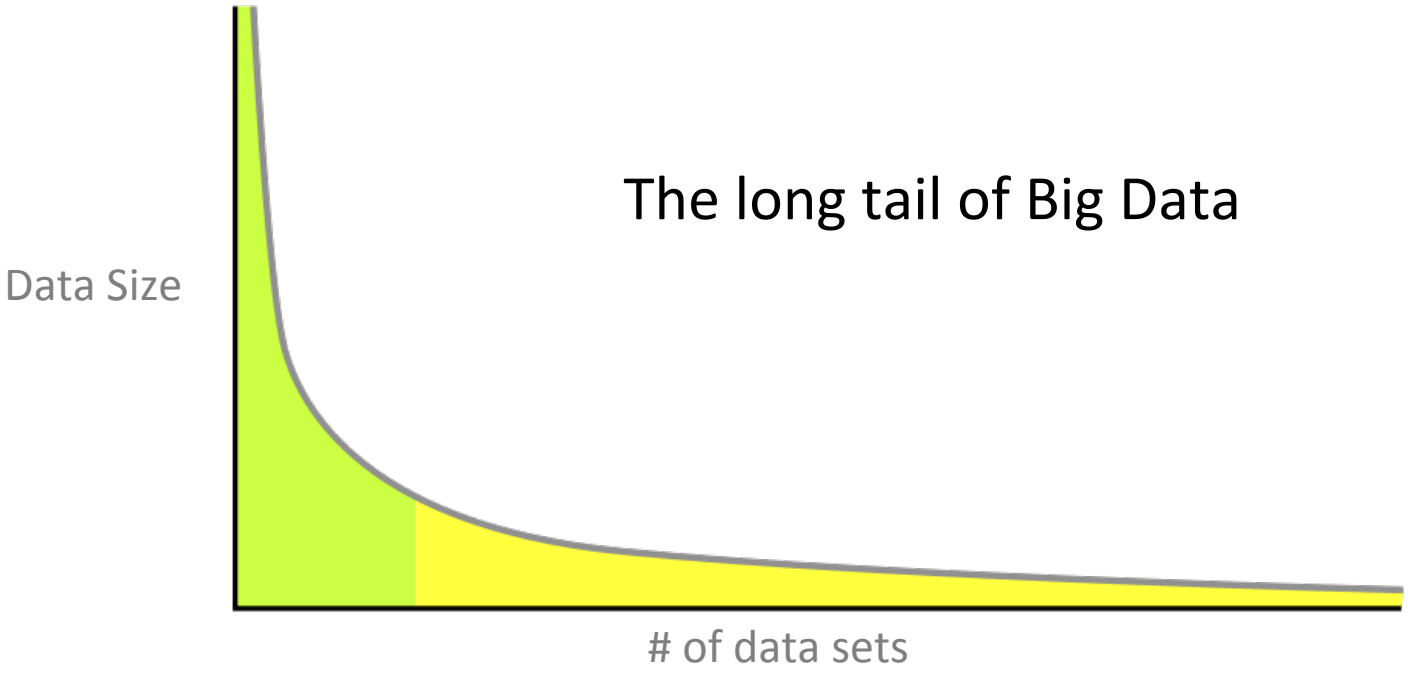
OVERVIEW

WHAT RESEARCHERS WANT AND WHAT RESEARCHERS DO

WHAT WE CAN DO BETTER: 10 SIMPLE RULES

INITIATIVES AND TOOLS SUPPORTING THE RULES

Thousands of scientists producing millions of data sets



- TB-PB scale data sets are often not at risk, managed by large data centers or organizations
- But the vast majority of data sets are smaller and managed (or **not** managed) by individual investigators

Rajasekar et al, 2103, Sociometric methods for relevancy analysis of Long Tail Science Data

Oct 19. 2103

The Economist | World politics | Business & finance | Economics | Science & technology | Cult

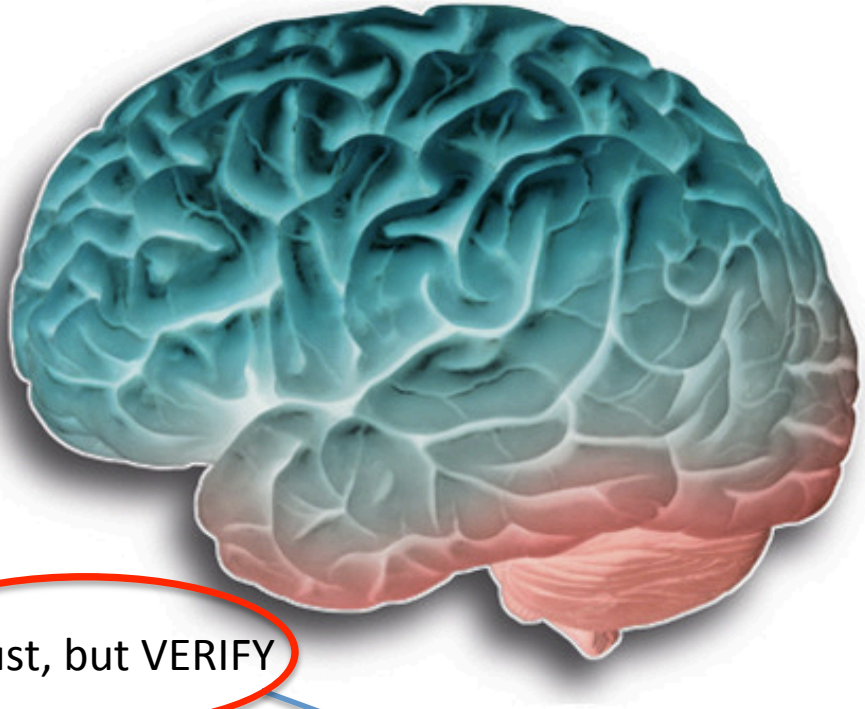
Problems with scientific research

How science goes wrong

Scientific research has changed the world. Now it needs to change itself

Oct 19th 2013 | From the print edition

"Ideally, research protocols should be registered in advance and monitored in virtual notebooks."



"Where possible, trial **data** also should be **open** for other researchers **to inspect and test**."

Trust, but VERIFY

A SIMPLE idea underpins science: "trust, but verify". Results should always be subject to challenge from experiment. That simple but powerful idea has generated a vast body of knowledge. Since its birth in the 17th century, modern science has changed the world beyond recognition, and overwhelmingly for the better.

Why “the care and feeding” of scientific data is important

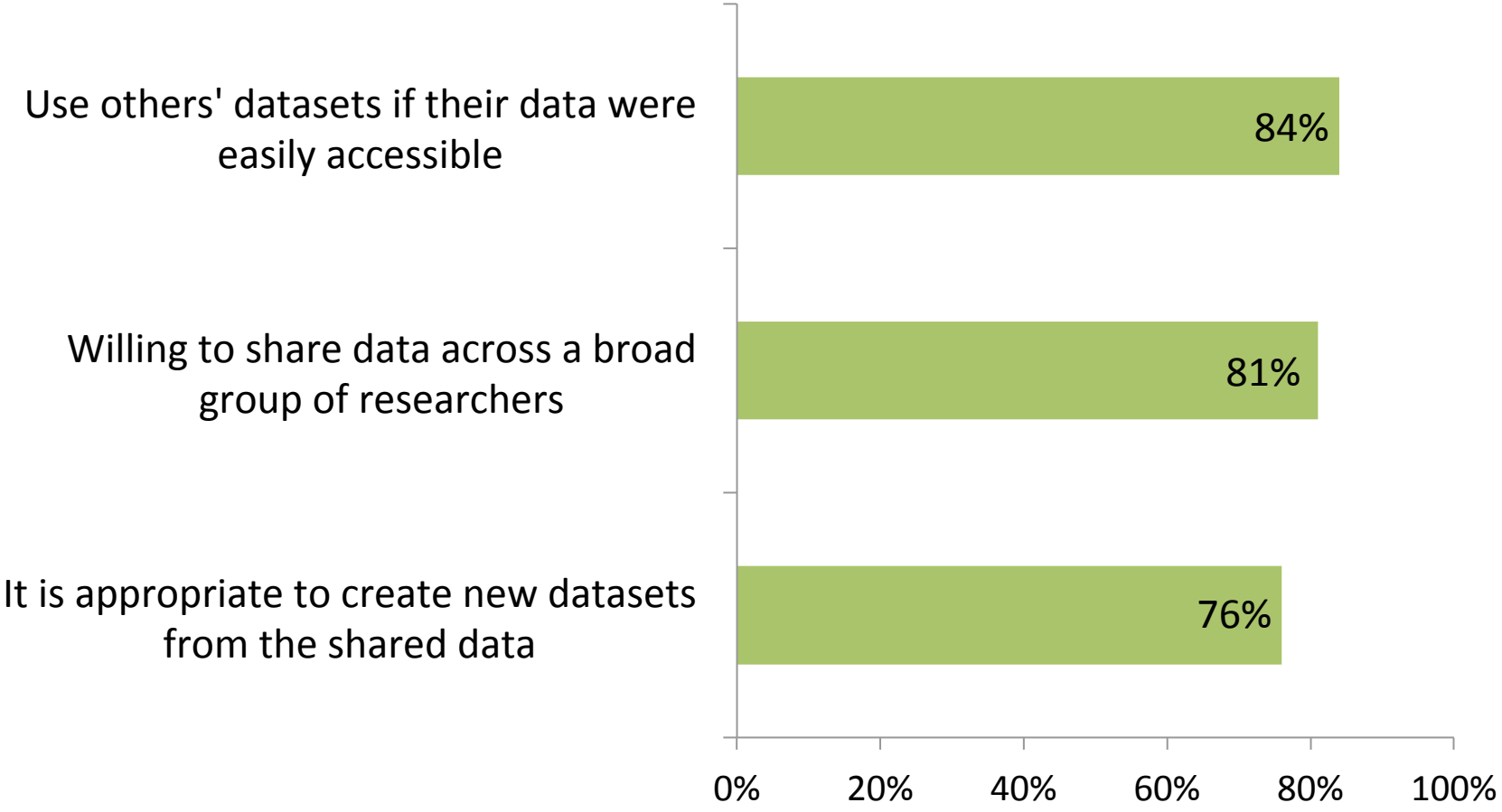
Christine L. Borgman, professor and Presidential Chair in Information Studies, UCLA, says:

- To reproduce research
- To make public assets available to the public
- To leverage investments in research data
- To advance research and innovation

Acknowledgement: Borgman, Oct 2013, “Why you should care about open data” Open Access Week Talk

Researchers *say* they want to share and reuse data

Online survey with 1315 respondents across disciplines (only 9% response rate); sample population mostly members of DataONE



Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, et al. (2011) Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 6(6): e21101. doi:10.1371/journal.pone.0021101 (Figure acknowledgement: Tenopir, U. of Tennessee)

46% made their data available in the Internet

36% agreed that their own data are easy to access

6% made all their data available

BUT DESPITE WILLINGNESS TO SHARE, ONLY A FEW DO

Data sharing is mostly demand-driven rather than supply-driven

Ten-year study with 22 random participants from the Center for Embedded Network Sensing (CENS)

"Data sharing tends to occur only through interpersonal exchanges."

"Ten of the 22 participants were unaware of repositories that would accept data from their type of research."

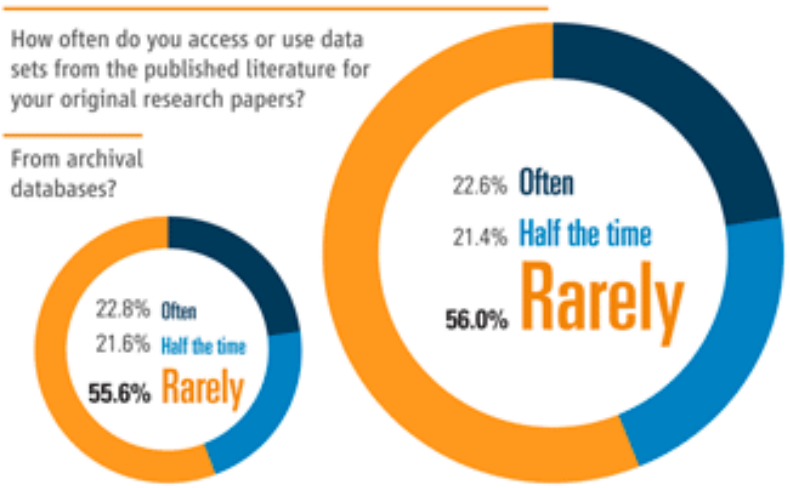
"14 participants said that they use data they themselves did not generate"

"Investigators want credit for their data, both in terms of first rights to publish their findings and in attribution for any reuse of their data."

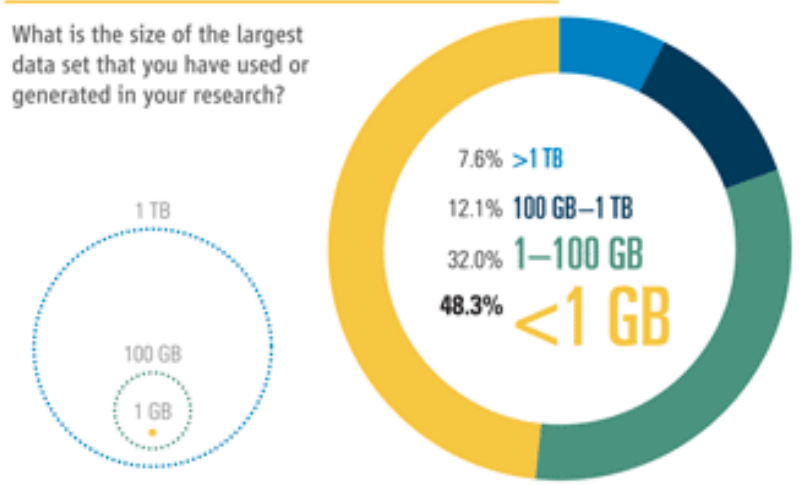
Wallis JC, Rolando E, Borgman CL (2013) If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. PLoS ONE 8(7): e67332. doi:10.1371/journal.pone.0067332

Survey with 1700 respondents from Science (2011) peer reviewers

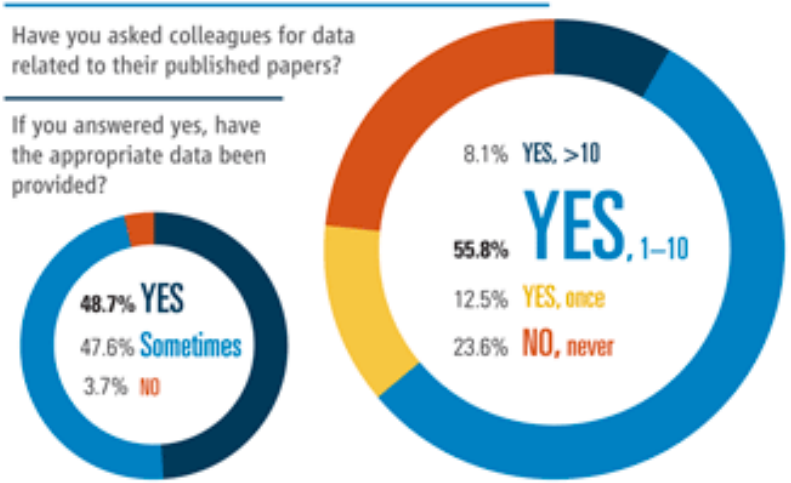
Access data from published work



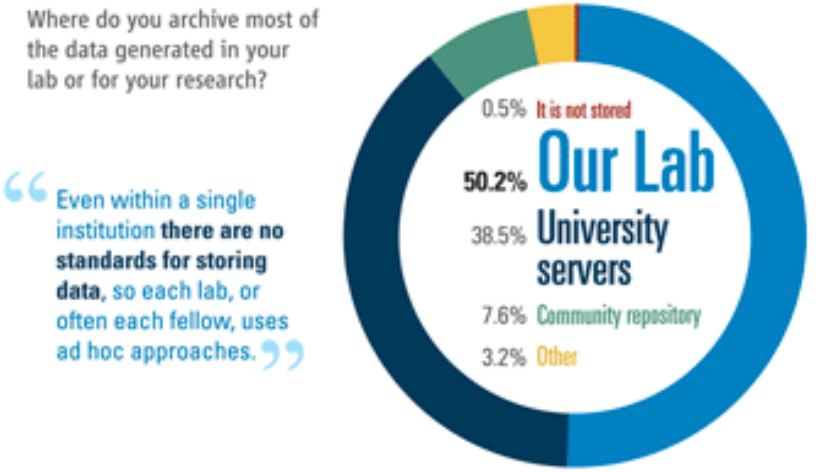
Size of data



Ask colleagues for data



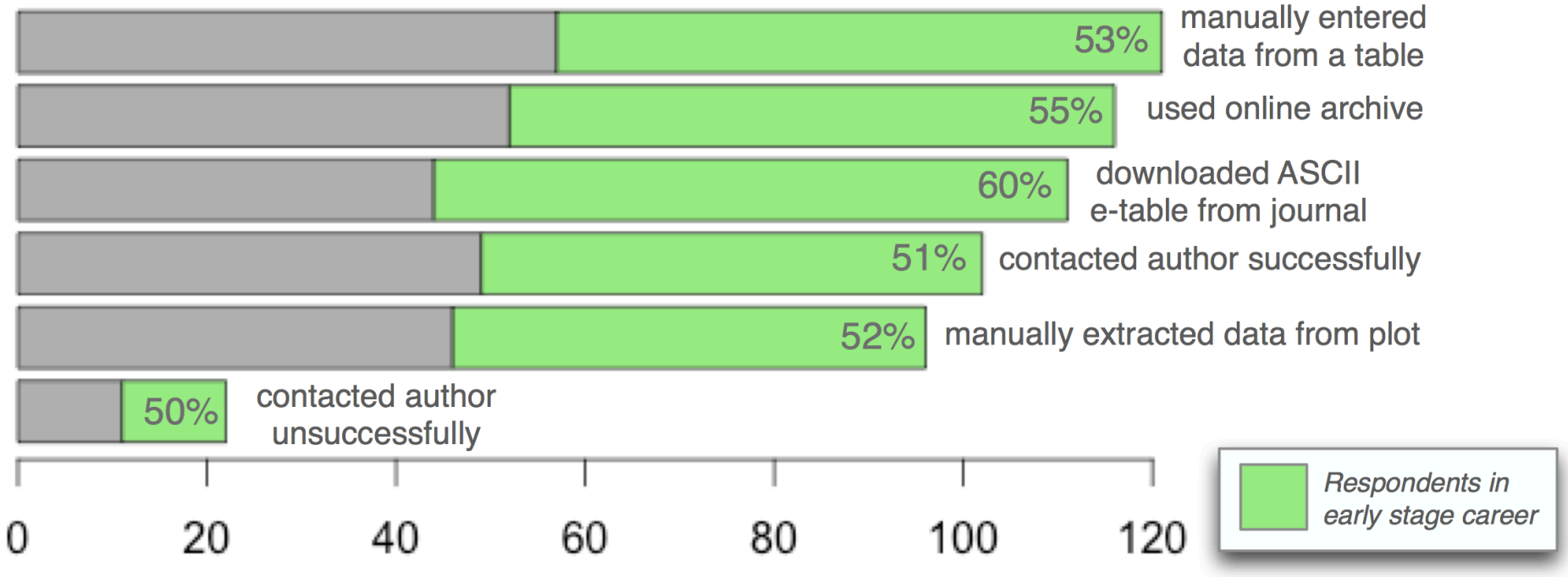
Archival location



Science 11 February 2011: Vol. 331 no. 6018 pp. 692-693 DOI: 10.1126/science.331.6018.692

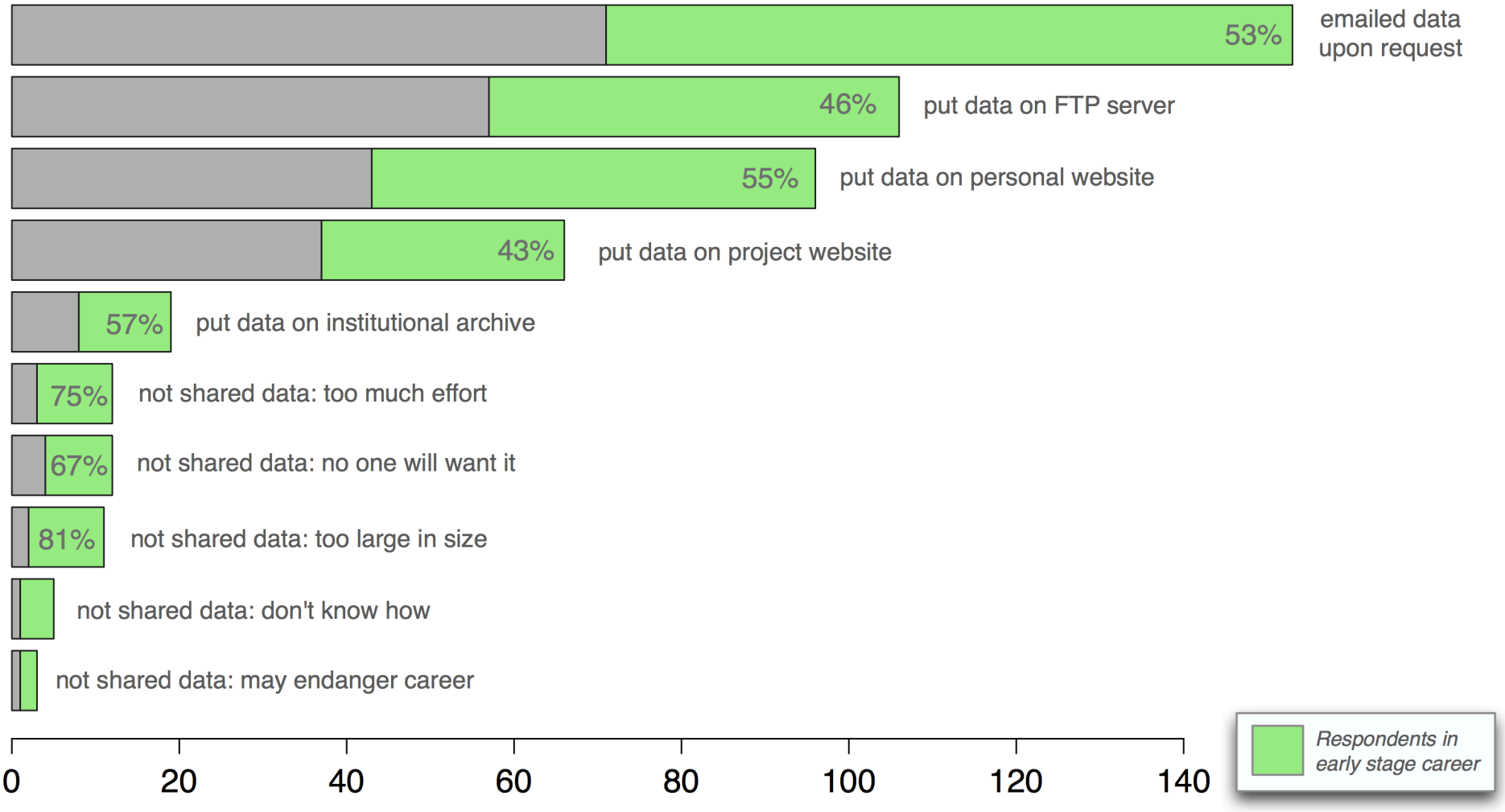
In Astronomy, a field with data standards

Survey sent to ~ 350 Ph.D. level researchers at the Harvard-Smithsonian Center for Astrophysics; 175 respondents



Have you ever used DATA you learned about from reading a Journal article?
Check ALL that apply

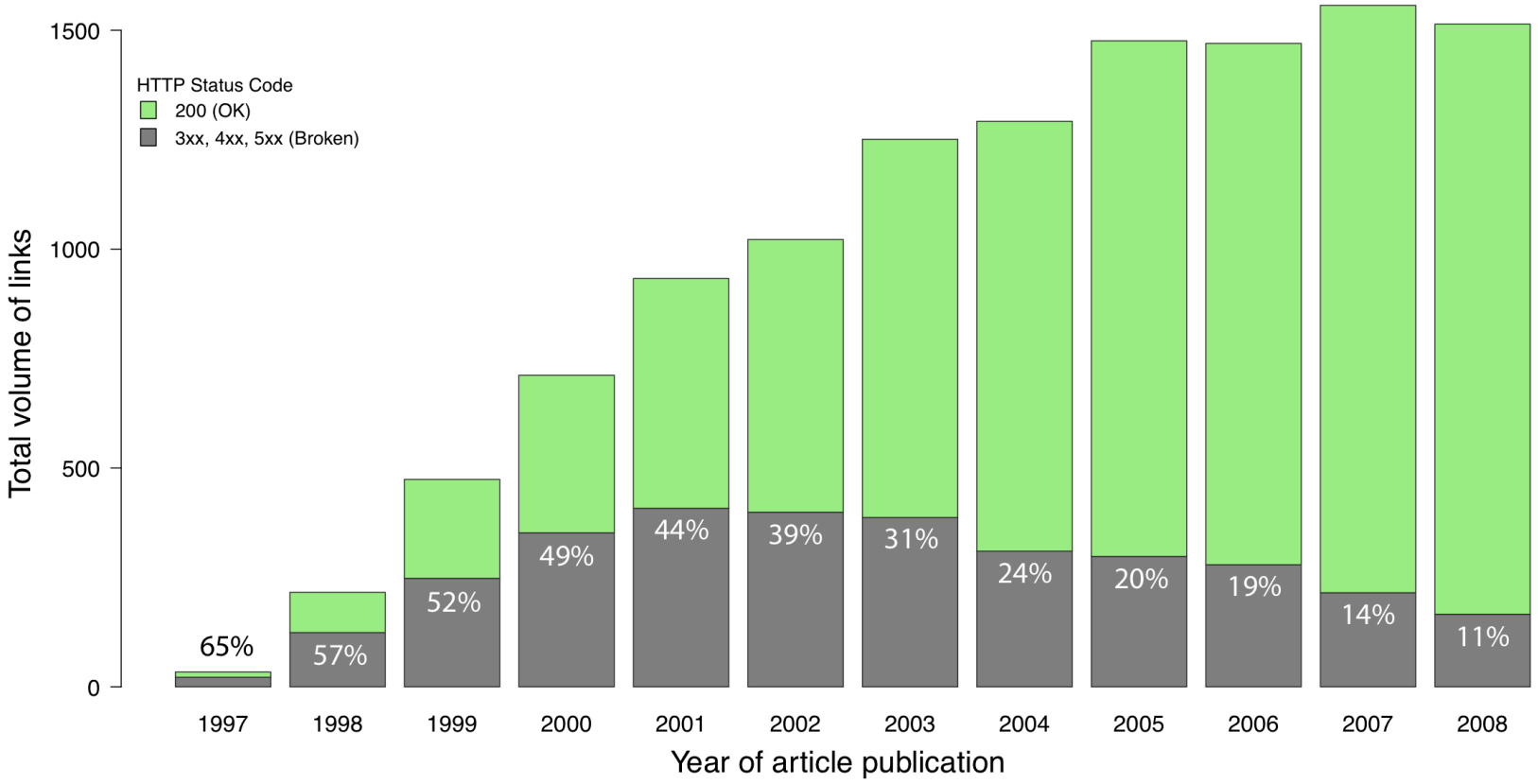
Pepe, Goodman, Muench, Crosas, Erdmann, 2013 "Sharing, Archiving and Citing Data in Astronomy" *Forthcoming*



When it comes to sharing DATA you've created, collected or curated, you have?
Check ALL that apply.

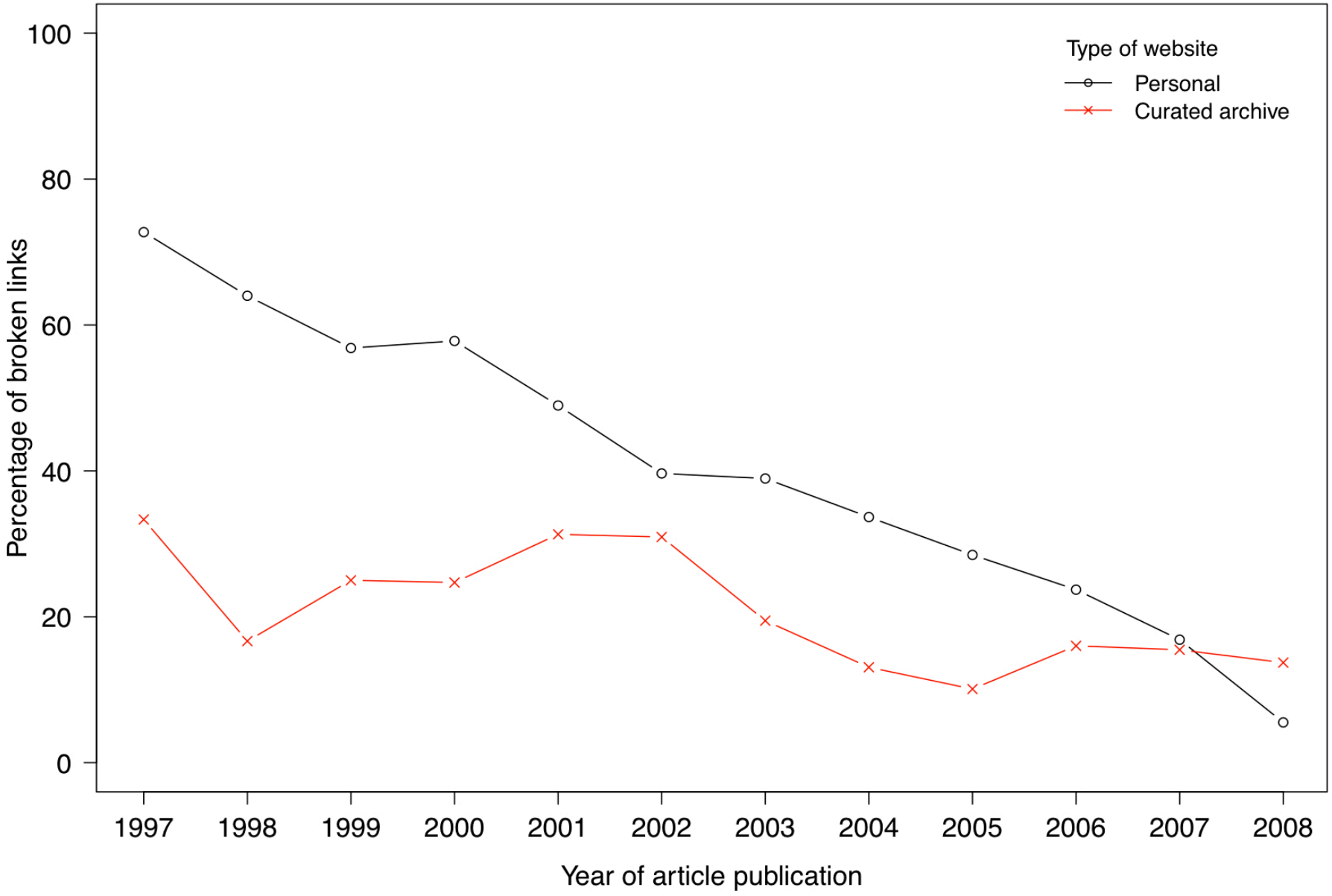
Pepe, Goodman, Muench, Crosas, Erdmann, 2013 "Sharing, Archiving and Citing Data in Astronomy" *Forthcoming*

Without persistent Data Citation, links broken over time



Links to data in astronomy publications between 1997 and 2008:
13,447 potential links to datasets in a total of 7,641 publication

Pepe, Goodman, Muench, Crosas, Erdmann, 2013 “Sharing, Archiving and Citing Data in Astronomy” *Forthcoming*



Pepe, Goodman, Muench, Crosas, Erdmann, 2013 “Sharing, Archiving and Citing Data in Astronomy” *Forthcoming*

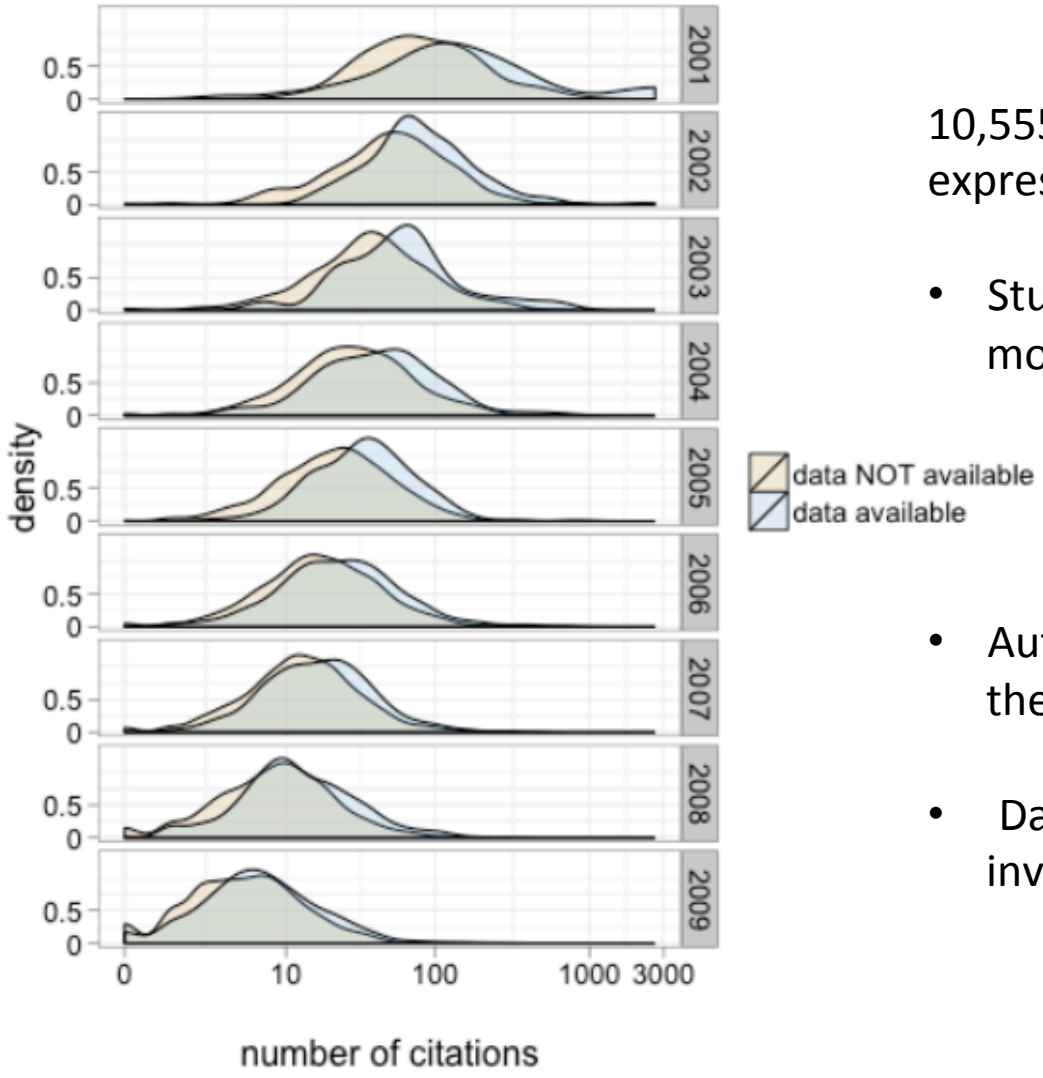
In Genomics, a field highly encouraged to share data

11,500 journal articles about **gene expression studies**, just **45 %** of the studies **shared their data**

Studies on **cancer** or studies with **human subjects, less** likely to **share data**

Piwowar and Vision (2013), Data reuse and the open data citation advantage. PeerJ 1:e175; DOI 10.7717/peerj.175

Studies that made data available get more citations



10,555 studies that created gene expression microarray data:

- Studies that share data received 9% more citations

- Authors published most papers using their own data within 2 years

- Data reuse paper by third-party investigators continued for 6 years

Piwowar and Vision (2013), Data reuse and the open data citation advantage. PeerJ 1:e175; DOI 10.7717/peerj.175

Rewards for publications

Competition, priority

Everyone is overwhelmed with life and email and, in academia, trying to get funding and write papers. Whether something is open or not open is not highest on the priority list. There's still need for making people aware of open science issues and making it easy for them to participate if they want to.

Effort to document data

Control, ownership

Jonathan Eisen, genetics professor at the University of California, Davis

DESPITE BEING GOOD FOR YOU AND FOR SCIENCE, TOO MANY CHALLENGES AND TOO LITTLE TIME

Schesinger (2013), "Scientist Threatened by Demands to Share Data", Aljazeera America

OVERVIEW

WHAT RESEARCHERS WANT AND WHAT RESEARCHERS DO

WHAT WE CAN DO BETTER: 10 SIMPLE RULES

INITIATIVES AND TOOLS SUPPORTING THE RULES

Make your data easily available to others, others are more likely to do the same—eventually.

At least take solace in the fact that you'll be able to find and reuse your own data if you treat them well.

RULE 1: LOVE YOUR DATA, AND LET OTHERS LOVE IT TOO

***Privacy concerns addressed later in this talk**

Your personal web site is unlikely to be a good option for long-term data storage.

Release your data in a general or domain-specific data archive

RULE 2: SHARE YOUR DATA ONLINE, WITH A PERMANENT IDENTIFIER

The higher the quality of provenance information, the higher the chance of enabling data reuse.

Keep: 1) data, 2) metadata, and 3) information about the process of generating those data, such as code.

RULE 3: CONDUCT SCIENCE WITH DATA REUSE IN MIND

Release a description of your processing steps to offer essential context for interpreting and re-using data.

RULE 4: PUBLISH WORKFLOW AS CONTEXT

Many journals now offer standard ways to contribute data to their archives and link it to your paper.

Use a formal data citation in the publication's reference list.

RULE 5: LINK YOUR DATA TO YOUR PUBLICATIONS AS EARLY AS POSSIBLE

Same best practices in relation to data and workflow also apply to software materials.

RULE 6: PUBLISH YOUR CODE

Simply describe your expectations on how you would like to be acknowledged.

You can also release your data under a license, but making it simple for others to reuse it, when possible.

RULE 7: SAY HOW YOU WANT TO GET CREDIT FOR YOUR DATA (AND SOFTWARE)

Seek help from librarians, archivists or research communities on domain-based repositories and generic repositories available.

RULE 8: FOSTER AND USE DATA REPOSITORIES

Praise those following good practices.

Follow good scientific practice and give credit to those whose data you use.

RULE 9: REWARD COLLEAGUES WHO SHARE THEIR DATA PROPERLY

Advocate for hiring data specialists and for the overall support of institutional programs that improve data sharing.

Teach whole courses, or mini-courses, related to caring for data and software, or incorporate the ideas into existing courses.

RULE 10: HELP ESTABLISH DATA SCIENCE AND DATA SCIENTIST AS VITAL

OVERVIEW

WHAT RESEARCHERS WANT AND WHAT RESEARCHERS DO

WHAT WE CAN DO BETTER: 10 SIMPLE RULES

INITIATIVES AND TOOLS SUPPORTING THE RULES

Initiatives from Funding Agencies

National Science Foundation (NSF), since 2011:

- "Investigators are expected to share with other researchers […] the primary data"
- "Proposal must include […] Data Management Plan"

National Institutes of Health (NIH), since 2003:

- "The NIH expects and support the timely release and sharing of final research data"

NASA:

- "Promotes the full and open sharing of all data"

U.S. Federal Policy, 2013:

- Open access to publications and data

- Open Data Initiatives:
  - "make public data available and accessible"

wellcome trust

BILL & MELINDA GATES foundation

Initiatives from Journals

Journal of open psychology data

ubiquity press open access

Recommended Repositories

The following repositories meet our peer-review requirements and are recommended for the archiving of JOPD datasets. Please contact usif you would like to use another repository or recommend that we add it to our list.

International repositories	National repositories	Institutional repositories
JOPD Dataverse	DANS	
Dryad	Gesis	
Figshare	FORS	
OpenfMRI	Odum	
Zenodo	SND (	
	TARKI	

F1000Research

Articles | For Authors | Blog | Advisory Panel | About / Contact

Peer review – credit where credit's due

Publishing small research units – interview with Ian Beales

Make your unpublished datasets work for you

POSTED BY VARSHA KHODIYAR, 27 AUGUST 2013

COMMENTS 0

In a recent Industry Forum report by Thomson Reuters, several pertinent statements were made:

- The growing accumulation of data produced by academics which is not destined for publication represents an impediment to scientific progress.
- Conventional research assessment methods do not recognise or reward data sharing.
- A researcher's overall contribution to scientific progress is greater than their peer-reviewed publication record.

At *F1000Research* we couldn't agree more!

Other Initiatives: **Data Citation Principles**

"Data should be considered legitimate, citable products of research."

- Principle 1: Importance
- Principle 2: Credit and Attribution
- Principle 3: Evidence
- Principle 4: Unique Identifiers
- Principle 5: Access
- Principle 6: Persistence
- Principle 7: Versioning and Granularity
- Principle 8: Interoperability and Flexibility

Endorsement by Publishers, Funding Agencies, Repositories and other stakeholders beginning in 2014

Synthesis group on Data Citation (includes Force11/Amsterdam Manifesto, CODATA, DataCite, UK Data Curation Center: http://www.force11.org/node/4381)

Data repositories

Examples of generic, self-curated data repositories:

Harvard Dataverse:
open, free to all

Free for public data,
Fees for non-public data

Open, free to all

Examples of domain-specific data repositories:

Biosciences; membership &
small fees for deposit

Pangaea: Earth and Environmental
Sciences; fees for some data submissions

NIH Genetic Sequence
Database

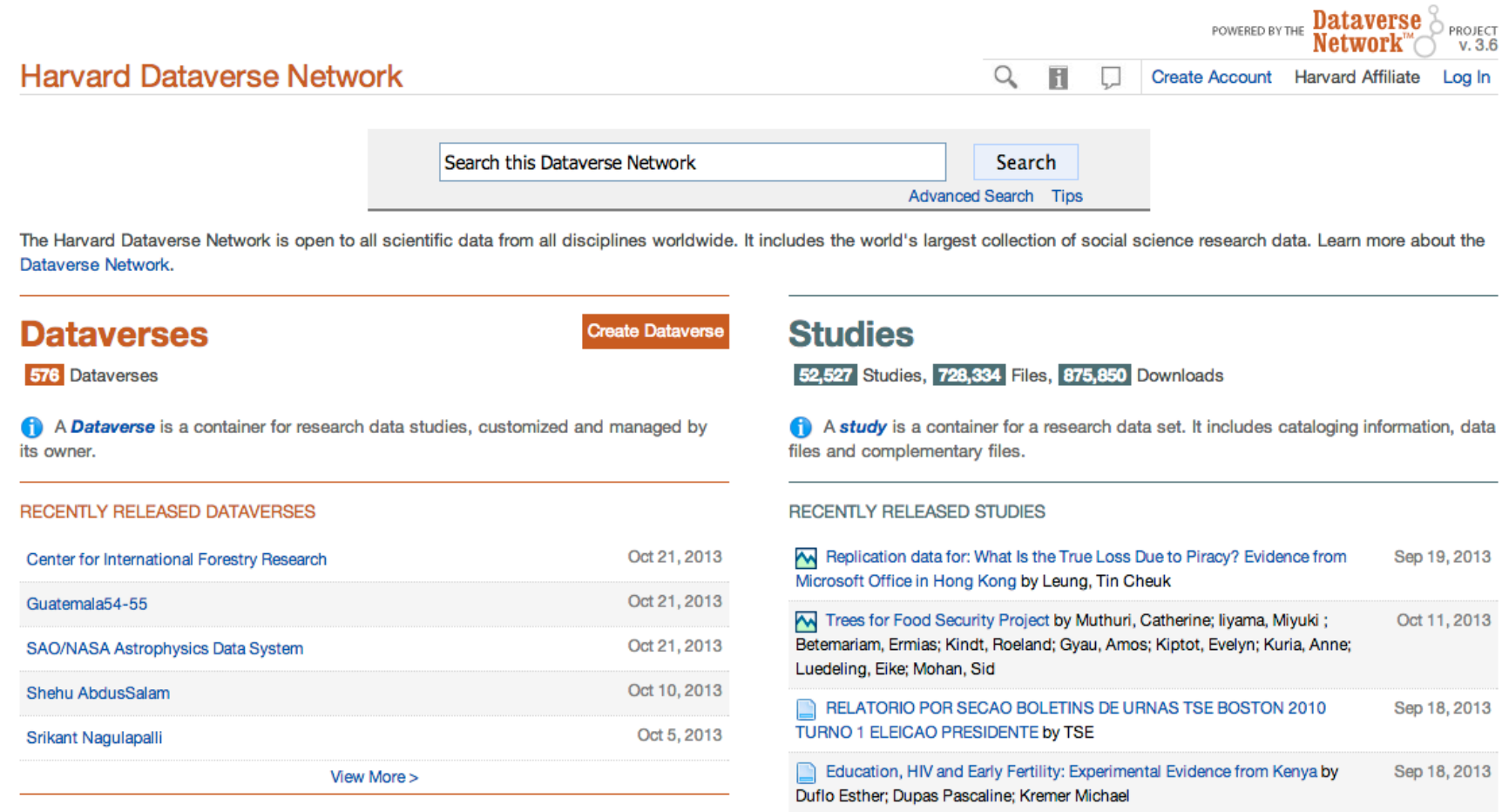
Examples of archives with curational services, and institutional repositories:

Data Curation for
Social Sciences

Sharing Data at Pen State

IQSS The Institute for Quantitative Social Science HARVARD UNIVERSITY | HARVARD LIBRARY | research DATA collaborative

Share, Cite, Reuse, Archive Research Data
Scientific data for reproducible research



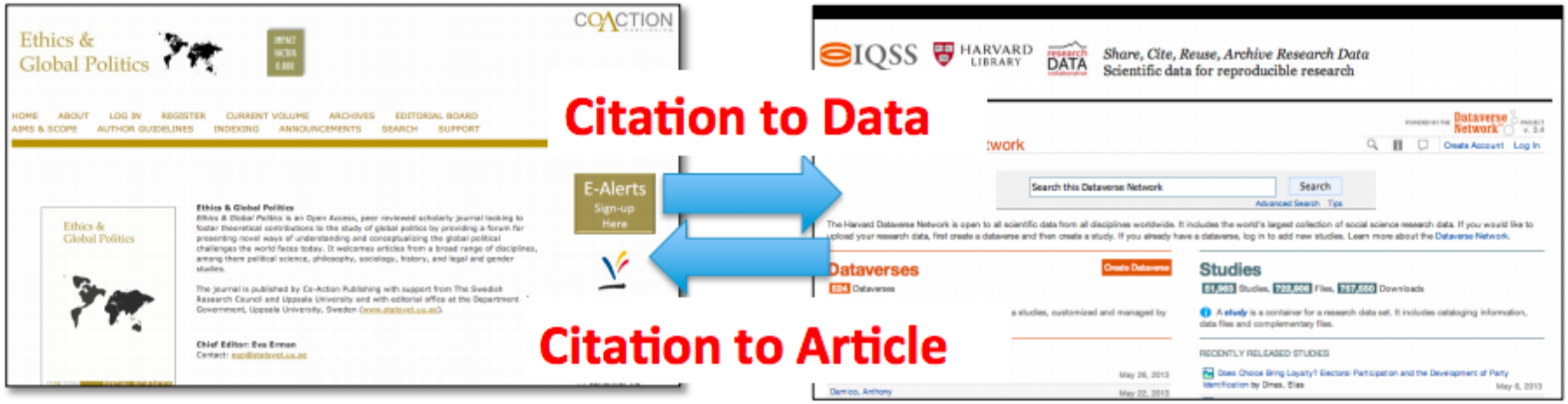
- Dataverse: open-source, data sharing & publishing framework, developed at IQSS
- Harvard Dataverse: Largest general purpose data repository

(> 52K studies with > 725K files)

Integration of Open Journal System (OJS) with Dataverse

OJS Journal

Harvard Dataverse Network



Phase I: 50 journals testing plugin

Phase II: 400 journals interested

Plugin:

Data + metadata + supporting files sent via SWORD API to the Dataverse

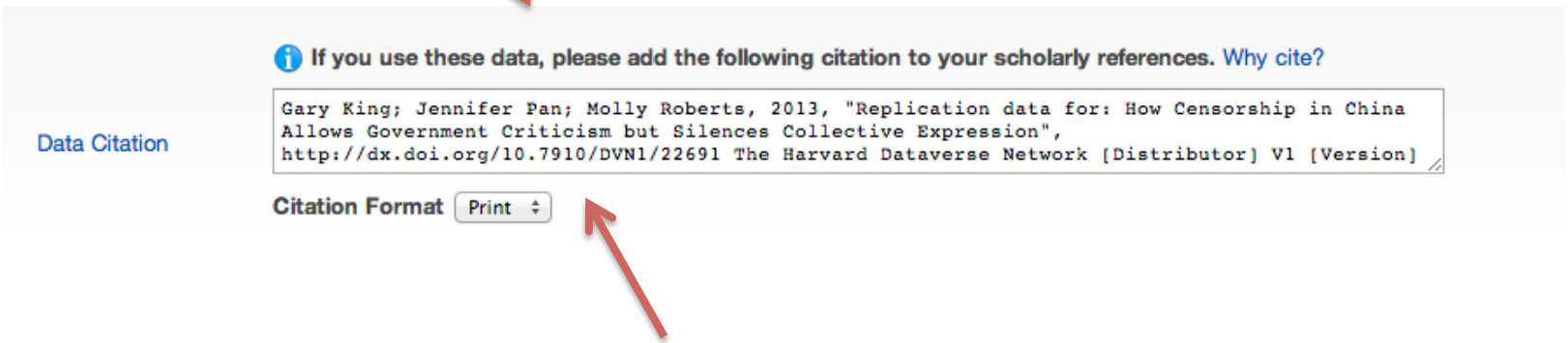
ALFRED P. SLOAN FOUNDATION 1934

PKP PUBLIC KNOWLEDGE PROJECT

The Dataverse Network™ Project

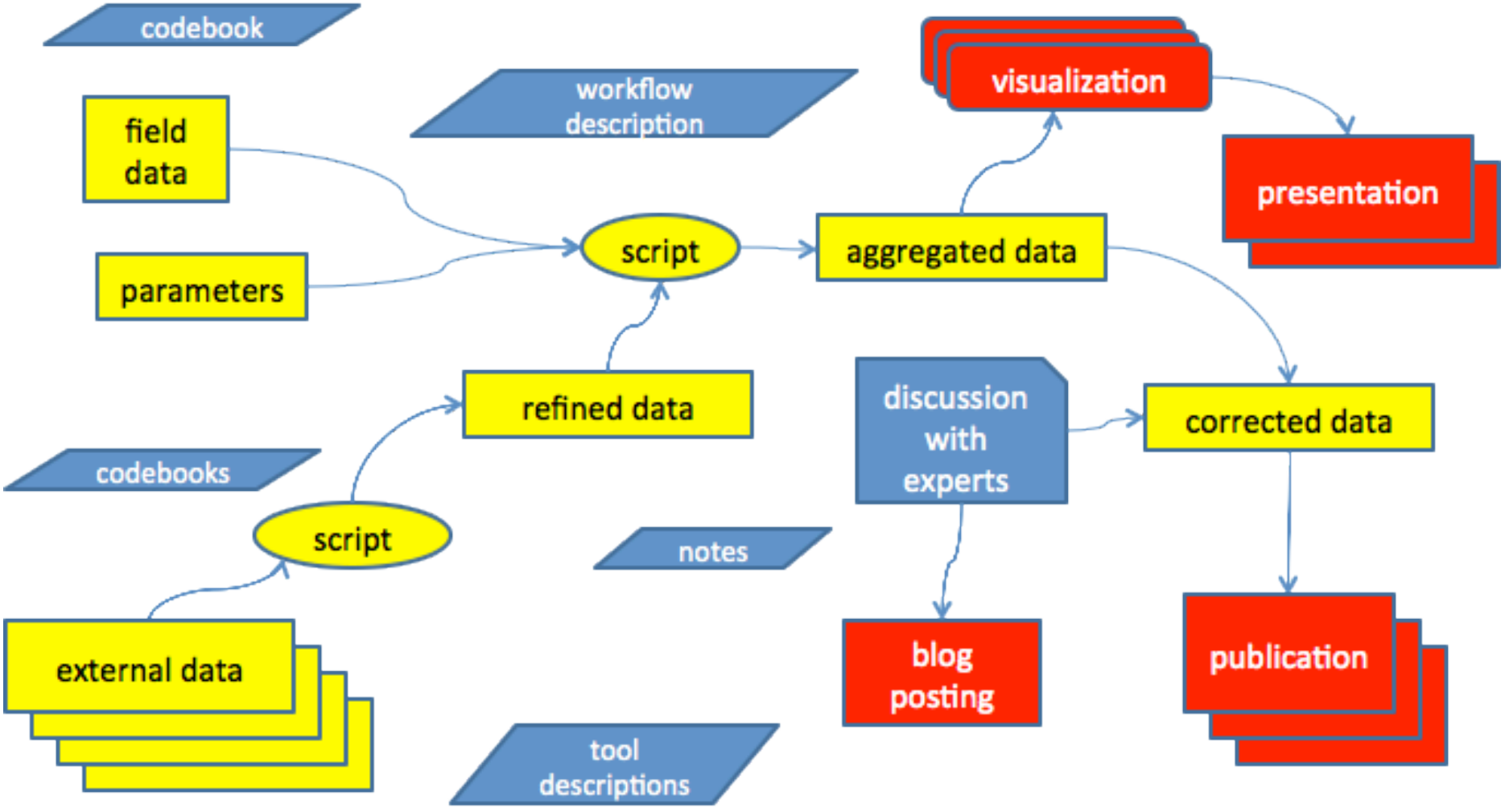
Formal Data Citation generated by Dataverse upon data set upload

Data Set can be referenced by using this Data Citation, giving credit to authors and repository



http://dx.doi.org/10.7910/DVN1/22691 resolves to the Data Set page, with description of the the data

Provenance and Workflows can be complex



Slide acknowledgment: Andrea Goethals

Workflow and Provenance Tools can help you

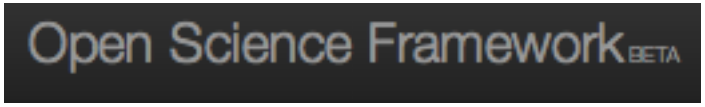
Domain Independent

For Biomedical Research

Tell stories with code and data

For Neuroscience Research

Share and execute code



Share, archive research materials, with data
Integrated with Dataverse, through SWORD API

Information Study Assay:
Rich description of experimental metadata

Data Formats in Social Sciences, Humanities and some Natural Sciences

Type of Data	Recommended Formats for Sharing, Reuse and Preservation
Tabular data, with extensive metadata	R, SPSS, Stata, SAS
Tabular data, with minimal metadata	CSV, tab-delimited, (xlsx)
Geospatial	ESRI shapefile, geo-referenced TIFF, CAD data, tabular GIS attribute data
Qualitative textual data	XML, rtf, txt
Digital image data	TIFF, (JPEG)
Digital audio data	FLAC, (MPEG-1, AIFF, WAV)
Digital video data	MPEG-4, motion JPEG 2000
Graph/social networks	GraphML
Documentation	rtf, PDF, ODT

Data Documentation Initiative (DDI): For generic tabular data metadata

Reference: UK Data Archive, http://data-archive.ac.uk/create-manage/format/formats-table

Formats used in specific Fields

Field	Recommended Formats for Sharing, Reuse and Preservation
Astronomy	Flexible Image Transport System (FITS)
High-Energy Physics	Based on ROOT (root.cern.ch) for PB data sets (for sharing outside – ASCII tabular data)
Genomics	FASTA, FASTQ, BAM, SAM, GEO, MUT, GCT, CEL, …
Flow Cytometry	FCS
Brain Imaging	DICOM, ….
Microscopy	OME-XML
Systems Biology	CELL-ML, BioPAX, SBML, SBGN, MIRIAM, …

+ Hundreds of OWL Vocabularies in NCBO BioPortal

For Biological data, increasing # of formats with increasing #of instruments!
Limited schema integration possible

Acknowledgment: Tim Clark (MGH, HMS), Kyle Cranmer (NYU)

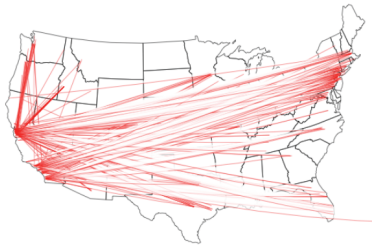
Let technology help you: Dataverse Features

SPSS, Stata, R Data

metadata extraction, subsetting & analysis (R, Zelig)

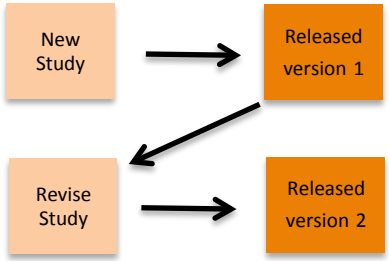
FITS Data

metadata extraction from file header



Social Network Data (GraphML)

smart queries & subsetting



Data Versioning

preserve & cite previous versions

Open Data Licenses and Badges

- Creative Commons:
  - CC0 for databases, data files
  - CC-by, reuse with attribution
- Open Data Commons:
  - Public domain for data and databases (PDDL)
  - Attribution for data/databases (ODC- by)
  - Attribution Share-Alike for Databases (ODBbL)
- Badges to acknowledge Open Practices:

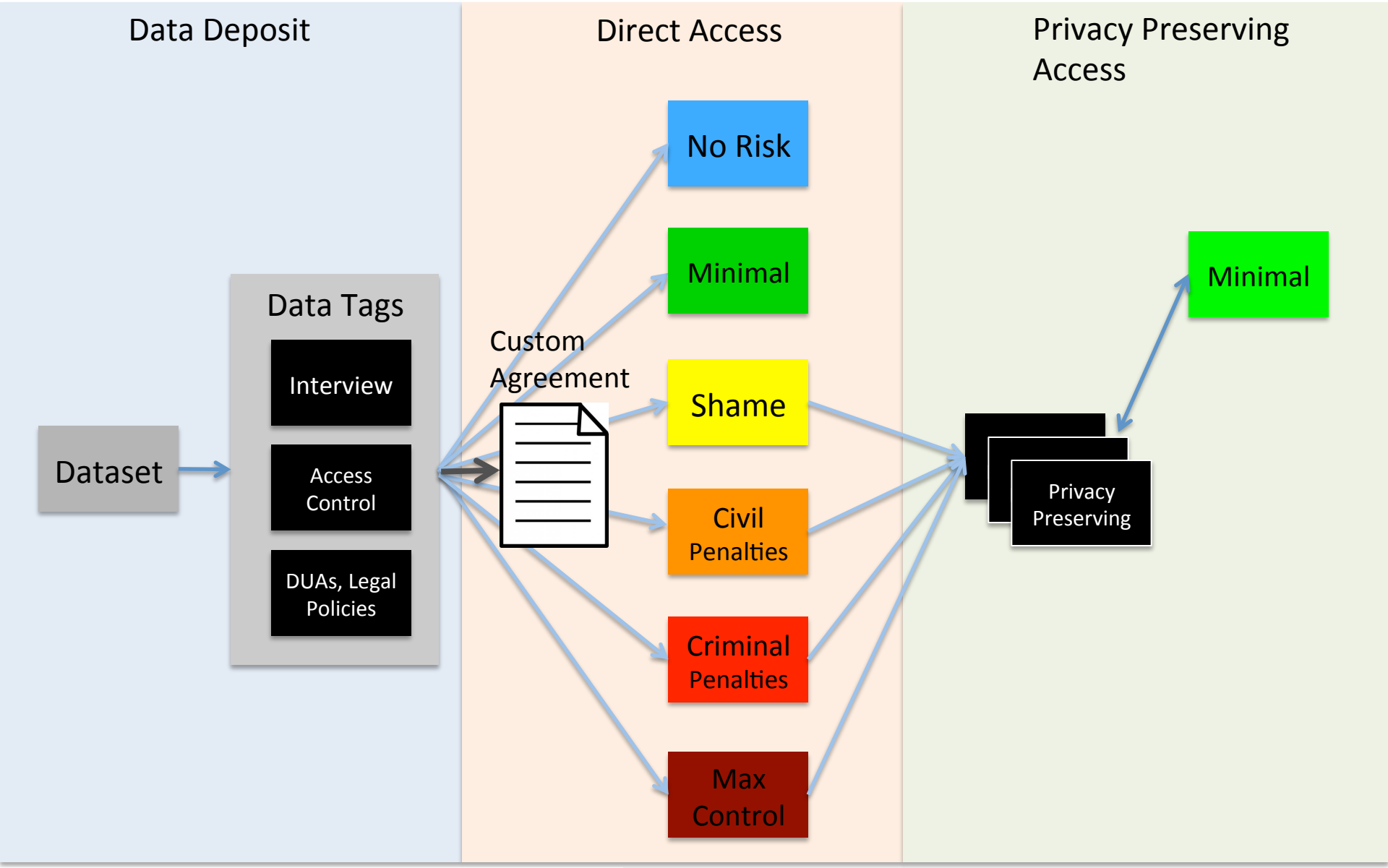
VIEW THE **BADGES:**

Open Science Framework BETA

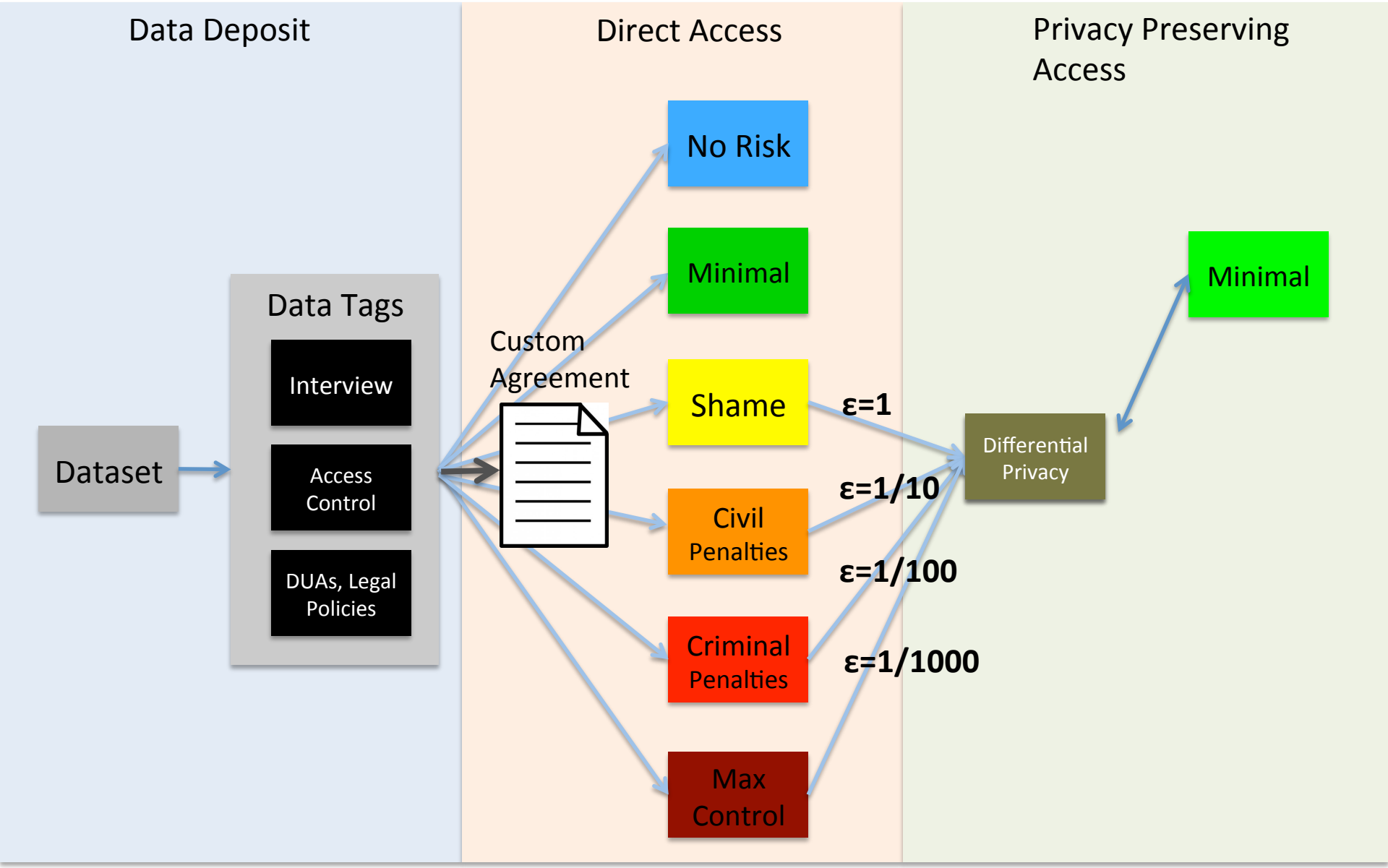
ClinicalTrials.gov

BUT NOT ALL DATA CAN BE 100% OPEN

Data Tags: Sharing Data with Confidence

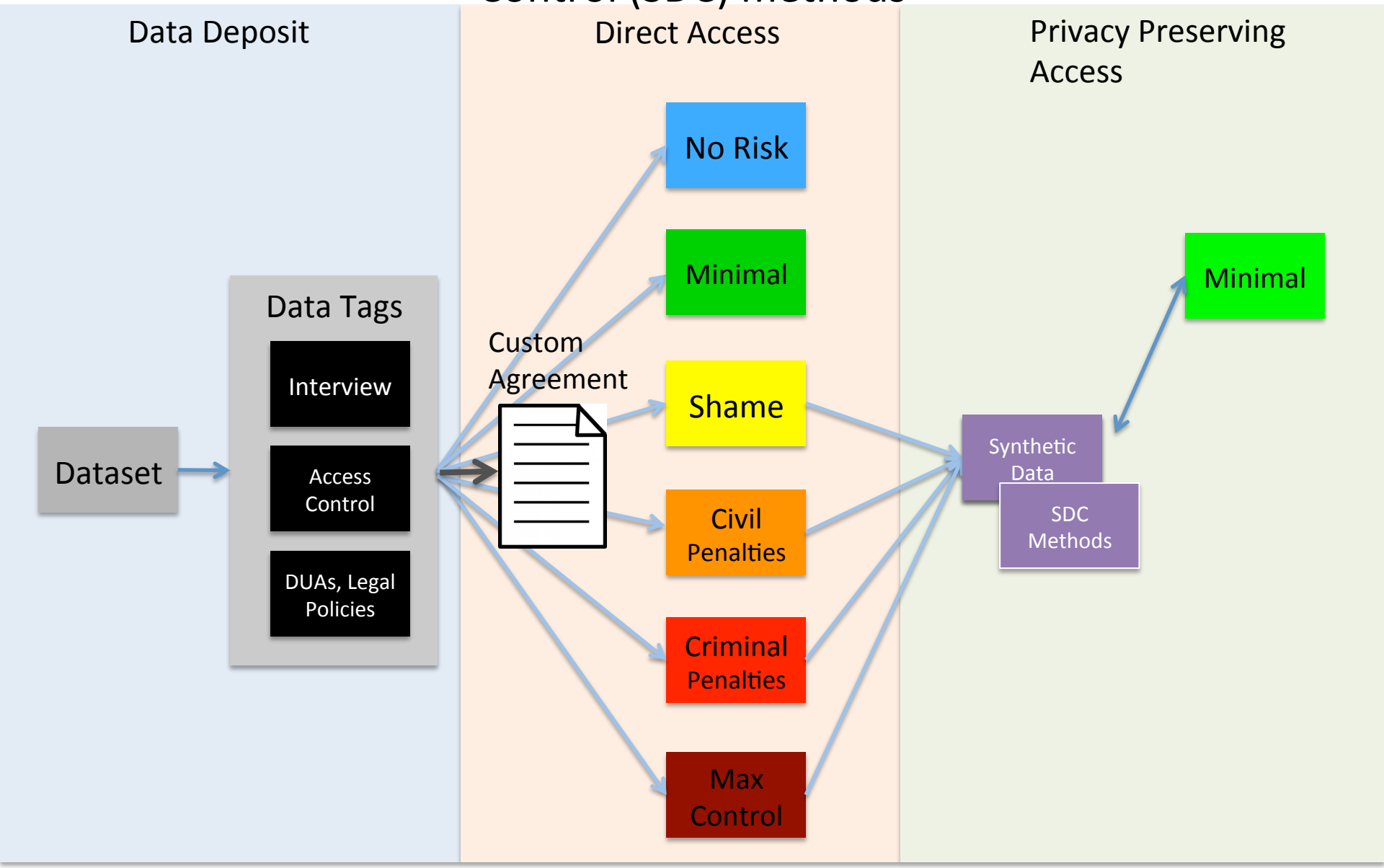


Privacy Preserving Tools: Differential Privacy



References on Differential Privacy: Ullman, Vadhan, Nissim; DP with focus on statistical utility: Vu & Slavkovic (2009)

Privacy Preserving Tools: Synthetic Data & other Statistical Disclosure Control (SDC) methods

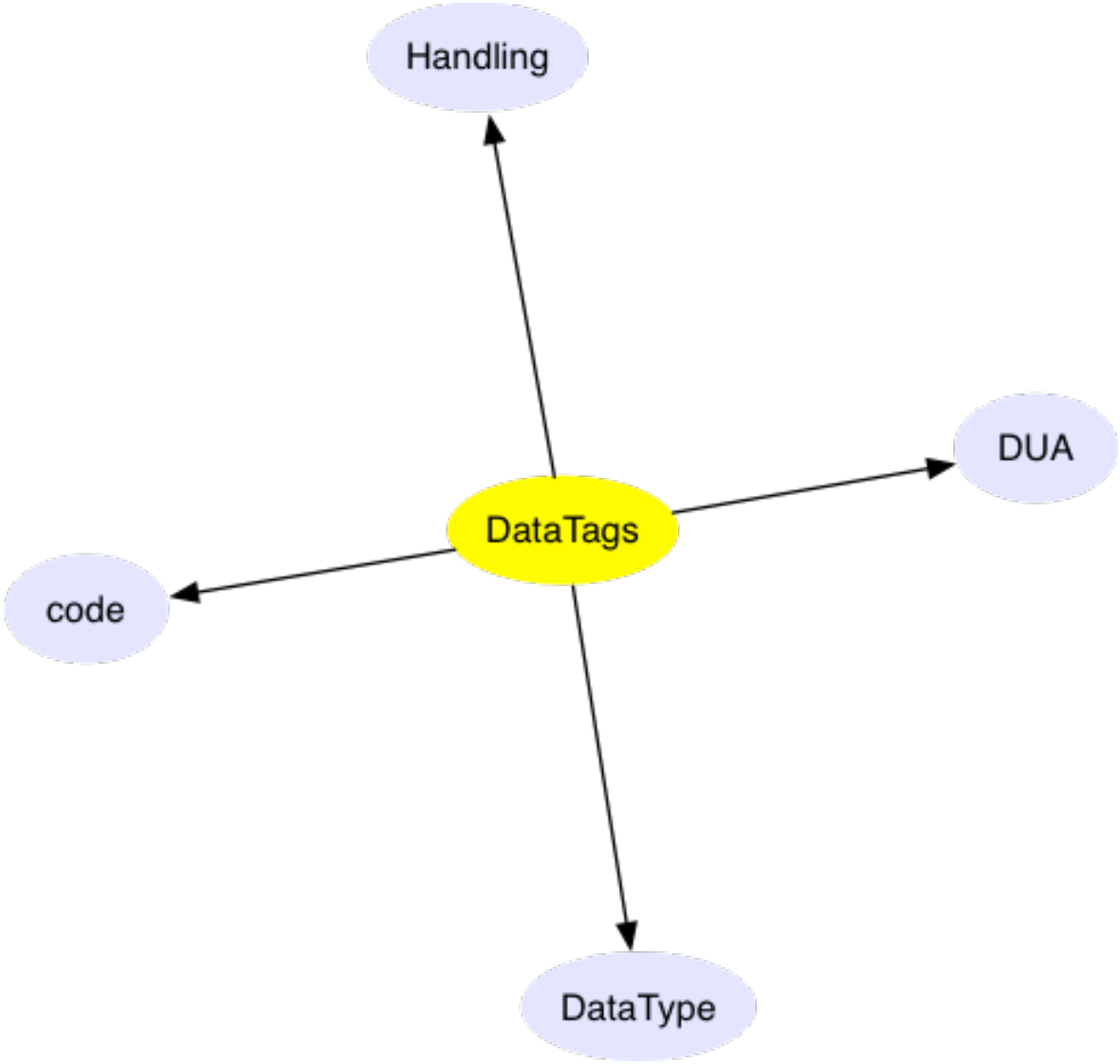


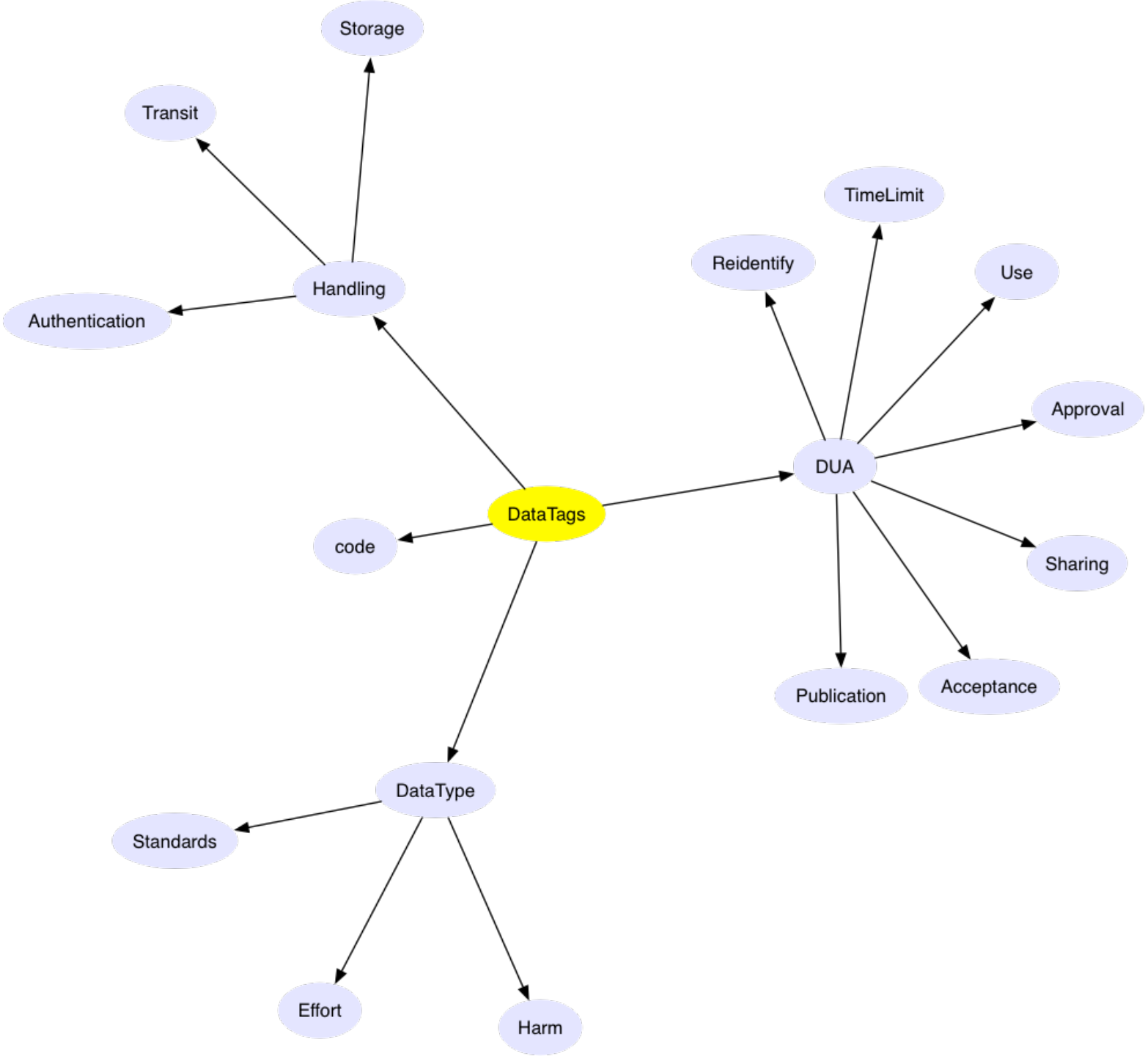
Reference: Slavkovic's talk on Privacy Preserving Methods

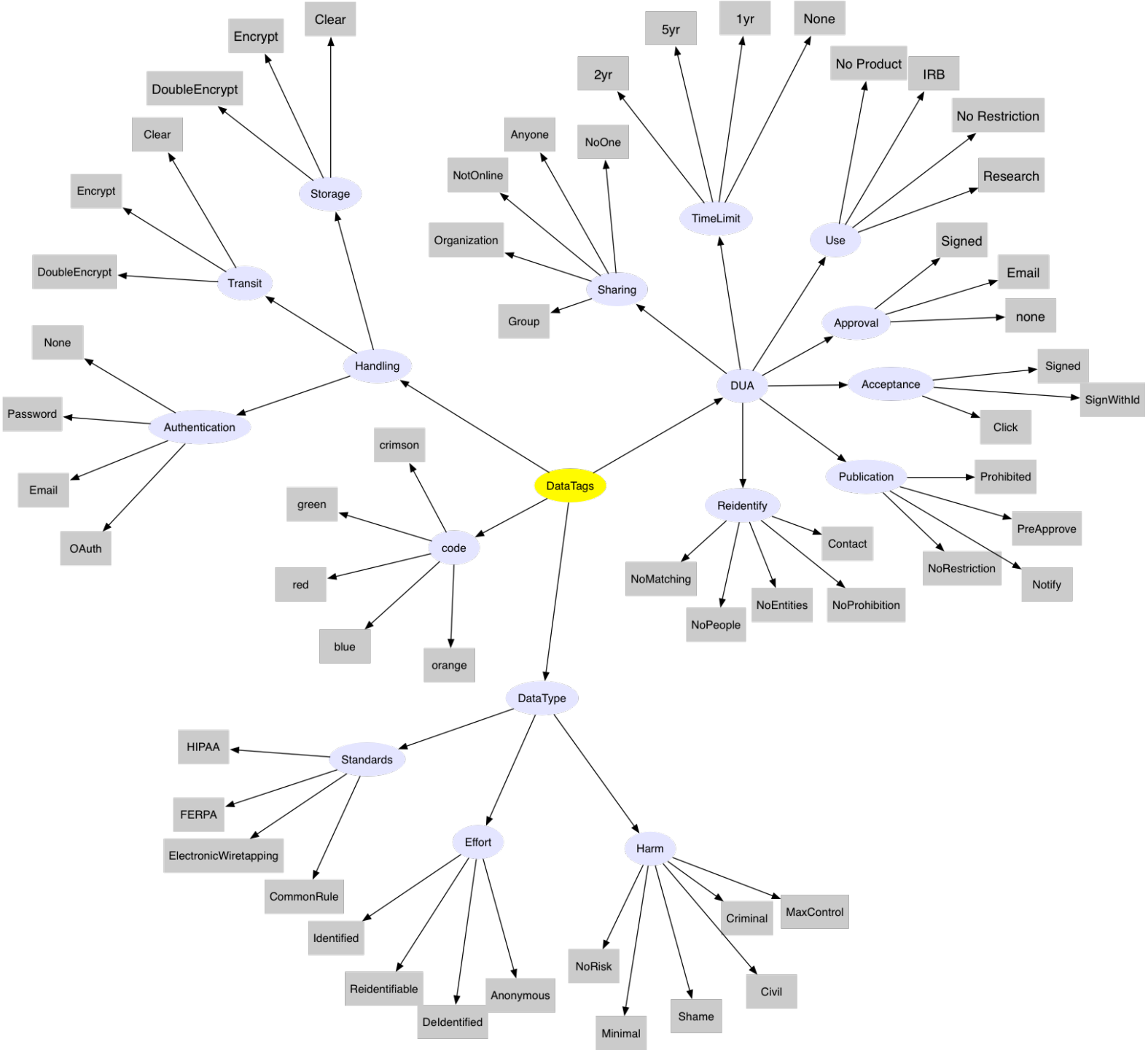
Data Tags Codes

- **BLUE:** **Non-confidential** information, stored and shared freely.
- **GREEN:** **Not harmful** personal information, shared with some access control.
- **YELLOW:** **Potentially harmful** personal information, shared with loosely verified and/or approved recipients.
- **ORANGE:** **Sensitive** personal information, shared with verified and/or approved recipients under agreement.
- **RED:** **Very sensitive** personal information, shared with strong verification of approved recipients under signed agreement.
- **CRIMSON:** **Maximum sensitive**, explicit permission for each transaction, strong verification of approved recipients under signed agreement.

Inspired by and compliant with Harvard Security Levels: http://security.harvard.edu/research-data-security-policy


Handling
DUA
DataTags
code
DataType


Storage
Transit
Handling
Authentication
TimeLimit
Reidentify
Use
Approval
DUA
DataTags
code
Sharing
Publication
Acceptance
DataType
Standards
Effort
Harm


DataTags
Handling
Storage
Clear
Encrypt
DoubleEncrypt
Transit
Clear
Encrypt
DoubleEncrypt
Authentication
None
Password
Email
OAuth
DUA
TimeLimit
5yr
1yr
None
Sharing
Anyone
NoOne
NotOnline
Organization
Group
Use
No Product
IRB
No Restriction
Research
Approval
Signed
Email
none
Acceptance
Signed
SignWithId
Click
Reidentify
NoMatching
NoPeople
NoEntities
NoProhibition
Contact
Publication
Prohibited
PreApprove
NoRestriction
Notify
code
crimson
green
red
blue
orange
DataType
Standards
HIPAA
FERPA
ElectronicWiretapping
CommonRule
Effort
Identified
Reidentifiable
DeIdentified
Anonymous
Harm
NoRisk
Minimal
Shame
Criminal
Civil
MaxControl

Privacy Laws

- 2187 privacy related laws in the United States, which can be grouped in ~ 30 types:

  Explicit consent, medicals records, student records, arrest and conviction records, bank and financial records, cable television, computer crime, credit reporting, criminal justice, electronic surveillance, employment records, government information on individuals, identity theft, insurance records (including use of Genetic information), library records, mailing lists, special medical records (including HIV testing), non-electronic surveillance, polygraphing in employment, privacy statutes/state constitutions, privileged communications, social security numbers, tax records, telephone services, testing in employment, tracking technologies, voter records.

Acknowledgement: Latanya Sweeney

Phase I Implementation

- DataTags.org web application, initially supporting:
  - Explicit Consent
  - Medical Records: HIPAA regulation
- Integrate with Secure **Dataverse**
- HIPAA certification
- Pilot at Harvard Medical School and School of Public Health

DATA TAGS DEMO

Acknowledgements: Data Science team at IQSS, Authorea, Latanya Sweeney, Michael Bar-Sinai, Christine Borgman, Tim Clark, Kyle Cramer, Aleksandra Slavkovic, Margaret Hestrom, Sonia Barbosa, Eleni Castro

THANK YOU