Research Data Management @Harvard

Mercè Crosas, Ph.D.

Harvard University's Research Data Officer, HUIT

Chief Data Science and Technology Officer, IQSS

Harvard Responsible Conduct of Research

August 14, 2019

"Research data management concerns the organization of data, from its entry to the research cycle through the dissemination and archiving of valuable results. It aims to ensure reliable verification of results, and permits new and innovative research built on existing information."

Whyte, A., Tedds, J. (2011). 'Making the Case for Research Data Management'. DCC Briefing Papers. Edinburgh: Digital Curation Centre

Why should you care about managing and sharing your research data?

- Helps *you* reuse your own data

- Facilitates reliable verification of results by others

- Permits new research built on existing data

- Fulfills data management and sharing plans required by federal funding agencies and other funders

- Lets you make public assets available to the public

- Allows you to publish datasets along with scholarly article, as now required by many leading journals

- Promotes use and citation of your work

- Helps you be compliant with University and Schools policies

Reproducibility has become a hot topic in the scientific community

Science Translational Medicine Home News Journals Topics Careers

Advancing precision medicine:
Current and future proteogenomic strategies for biomarker discovery and development

Download this booklet today!

Science AAAS
Sponsored by Olink Proteomics

Log in | My acc

SHARE

IN THE PIPELINE

Derek Lowe's commentary on drug discovery and the pharma industry. An editorially independent blog from the publishers of *Science Translational Medicine*. All content is Derek's own, and he does not in any way speak for his employer.

By Derek Lowe

Reproducibility: Crisis or Not?

By Derek Lowe | May 26, 2016

oducibility in the scientific literature. They
ot", but it shows that we're still arguing about what
s (over 1500 scientists) said that there was "a
eople's minds.

nature
International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For

Archive > Volume 533 > Issue 7604 > Editorial > Article

NATURE | EDITORIAL

Reality check on reproducibility

A survey of *Nature* readers revealed a high level of concern about the problem of

NIH National Institutes of Health
Turning Discovery Into Health

Search NIH

NIH Employee Intranet | Staff Directory | En Español

Health Information | Grants & Funding | News & Events | Research & Training | Institutes at NIH | About NIH

Home » Research & Training

RIGOR AND REPRODUCIBILITY

Rigor and Reproducibility

Reporting Guidelines
Application Instructions
Training
Funding Opportunities
Meetings and Workshops
Announcements
Publications
Resources

Two of the cornerstones of science advancement are rigor in designing and performing scientific research and the ability to reproduce biomedical research findings. The application of rigor ensures robust and unbiased experimental design, methodology, analysis, interpretation, and reporting of results. When a result can be reproduced by multiple scientists, it validates the original results and readiness to progress to the next phase of research. This is especially important for clinical trials in humans, which are built on studies that have demonstrated a particular effect or outcome.

Johns Hopkins University students in a laboratory. *Johns Hopkins University*

Email Updates

Sign up to receive email updates about rigor and reproducibility.

Sign up for updates

Contact Us

Please send email to NIHReprodEfforts@od.nih.gov

A comprehensive Report on **Reproducibility and Replicability in Science** was recently published by the National Academies of Sciences, Engineering, and Medicine


The National Academies of
SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

Reproducibility and Replicability in Science
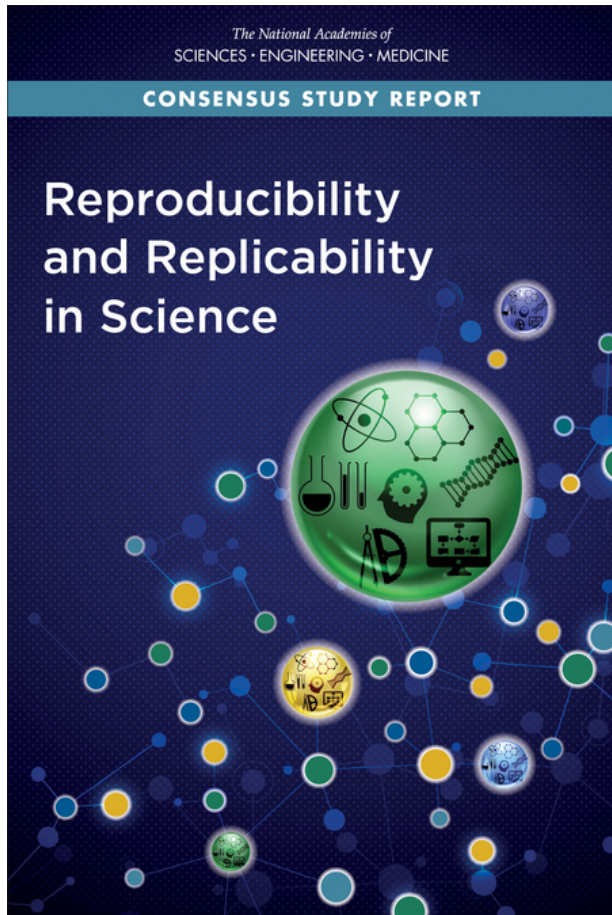
Report Highlights:

- **No crisis**, but we must do better.
- Include a clear, specific, and complete description of how results are reached:
    - all methods, instruments, materials, procedures;
    - the analysis of data and decisions for exclusion of some data or inclusion of other;
    - the analytic decisions and when these decisions were made and whether the study is exploratory or confirmatory;
    - a discussion of the expected constraints on generality
    - reporting of precision or statistical power; and
    - discussion of the uncertainty of the measurements, results, and inferences.
- Promote use of open source tools
- Journals should ensure computational reproducibility
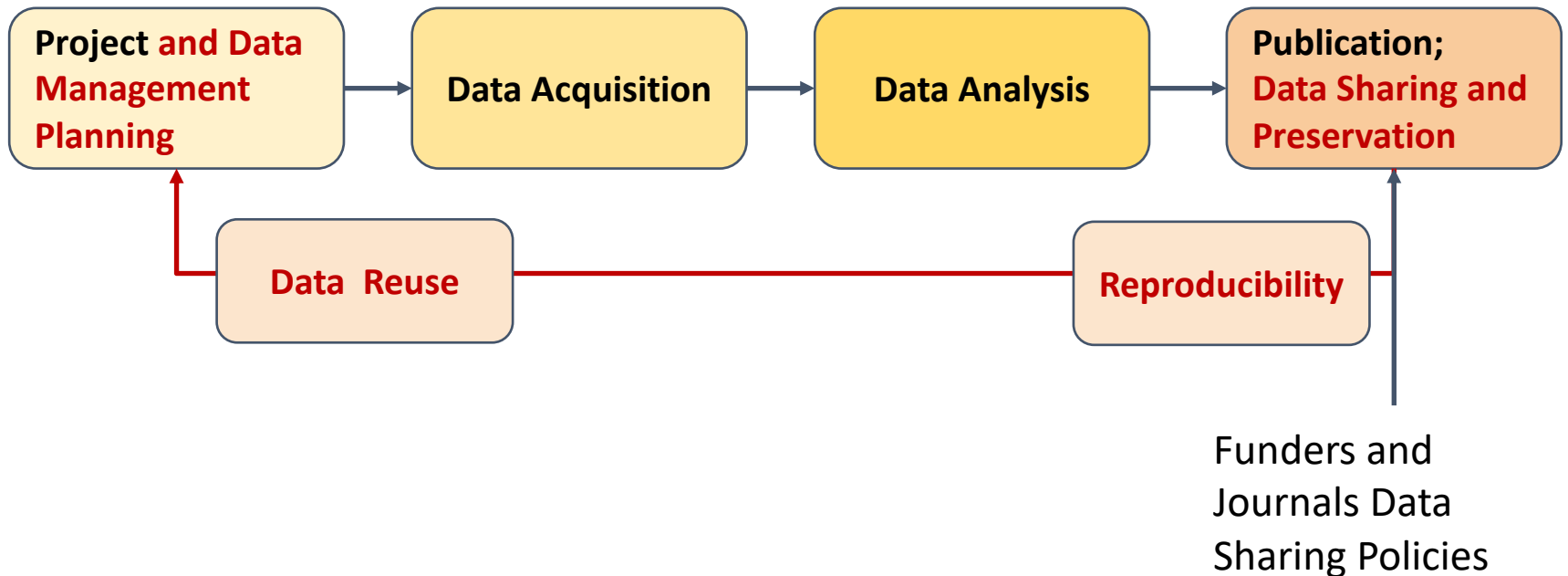- Facilitate transparent sharing and availability of digital artifacts, such as data and code

http://sites.nationalacademies.org/sites/reproducibility-in-science/index.htm

The Old Data Lifecycle
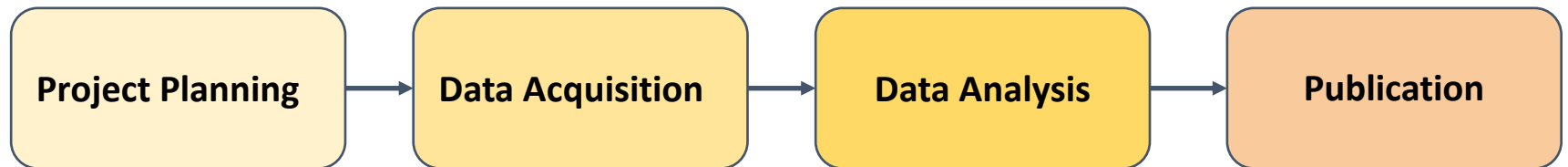
Data Roadmap, based on Briney, 2015. Data Management for Researchers

The New Data Lifecycle
Data as a Product of Research

Data Roadmap, based on Briney, 2015. Data Management for Researchers

New site and resources at Harvard support the new data lifecycle:
https://researchdatamanagement.harvard.edu

Harvard Library provides user support across the research data lifecycle

Individual libraries offer specialized data and digital scholarship services & resources

PLAN	COLLECT, GENERATE & STORE	CLEAN, ANALYZE & VISUALIZE	PUBLISH & SHARE	ARCHIVE & PRESERVE	REUSE
Plan for research data needs	**Acquire, organize & store data**	**Process data for current use**	**Organize & share data in repository**	**Appraise & steward data**	**Discover & reuse data**
DMPTool	Digitization services	Data cleaning, processing & visualization services & support	Harvard Dataverse data repository & DASH	Digital & data preservation services	Data reference services
Data Management Plans (DMPs	Data acquisitions		Data curation services	Consultations, referrals & best practices	Consultations & referrals
Consultations, best practices & referrals	Consultations & best practices	Consultations, referrals & best practices	Consultations & referrals		

Research Data Management is a collaborative effort: *You need to do your part*
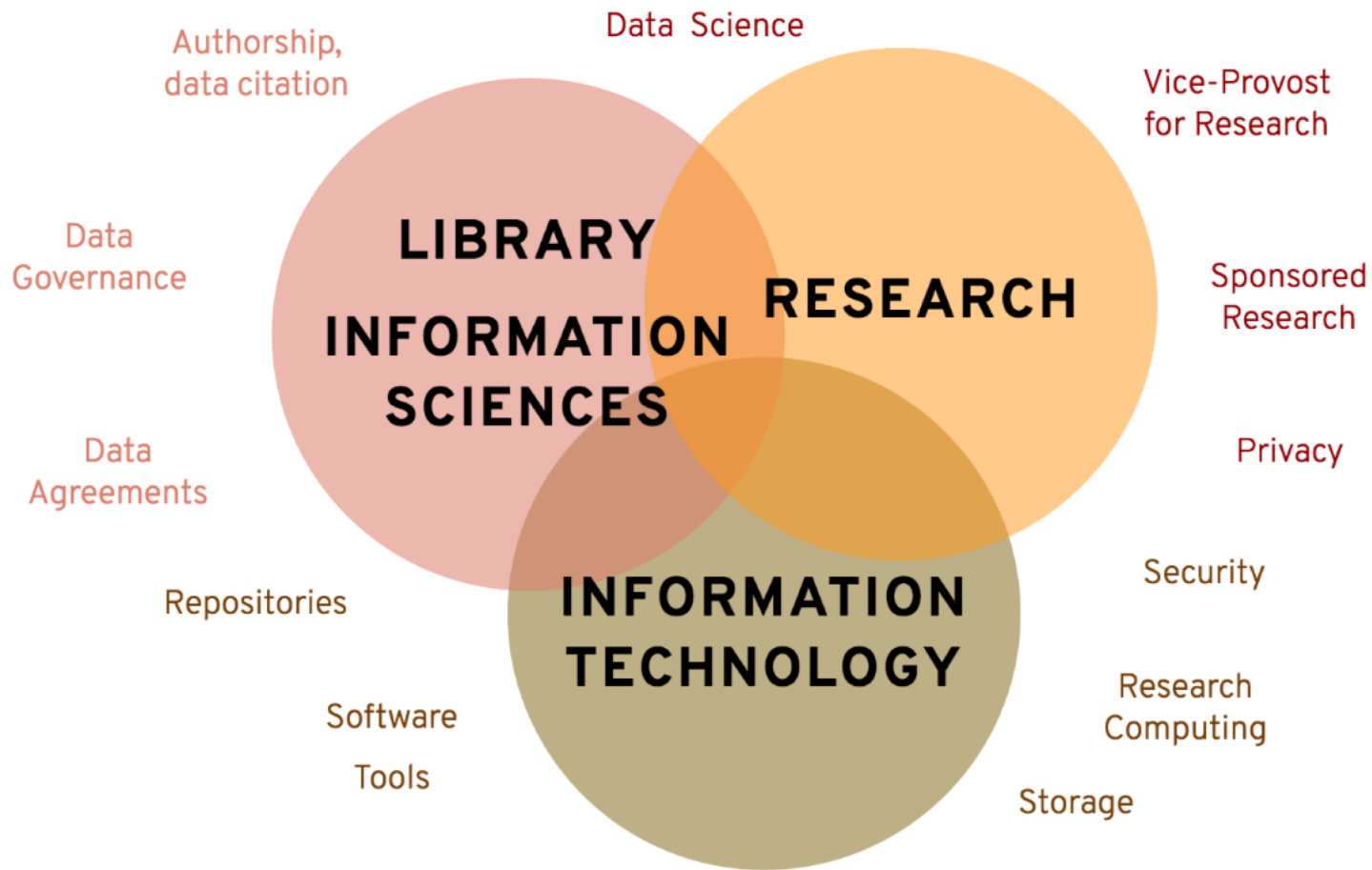
Planning for Data Management

| Planning | → | Data Acquisition & Collection | → | Storage, Security, & Analysis | → | Dissemination & Preservation |

Data Management Plans – Funders Requirements

- Federal: NSF, NIH, and most federal agencies require public access plans, data management and sharing plans

- Foundations: Sloan, Gates, Wellcome Trust, and most philanthropic organizations require some type of public access plans

- Your Data Management Plan (DMP) must be created and submitted as part of the funding application process

- **Plan ahead!**

| Planning | → | Data Acquisition & Collection | → | Storage, Security, & Analysis | → | Dissemination & Preservation |

DMPTool and DMP Support at Harvard

DMPTool:

- Online tool to help you create a Data Management Plan (DMP)
- Helps meet funder requirements and best practices
- Includes templates
- https://dmptool.org/ - Login with HarvardKey

Harvard Library Research Data Management program:

- Provides help to prepare your DMP
- https://hlrdm.library.harvard.edu/dmptool

Planning		Data Acquisition & Collection		Storage, Security, & Analysis		Dissemination & Preservation

Data Management Checklist

Should include:

- Type of data: observation, experimental, simulation, derived

- Form of data: quantitative/ qualitative, structure/unstructured

- File format: community standards, open formats (not proprietary)

- Size of data: fixed/streaming

- Metadata: community standards, when possible

- Code: if accompanies the data

- Documentation

- Plan where your data will be stored during your research (active phase), and where will be shared and archived after the research is published

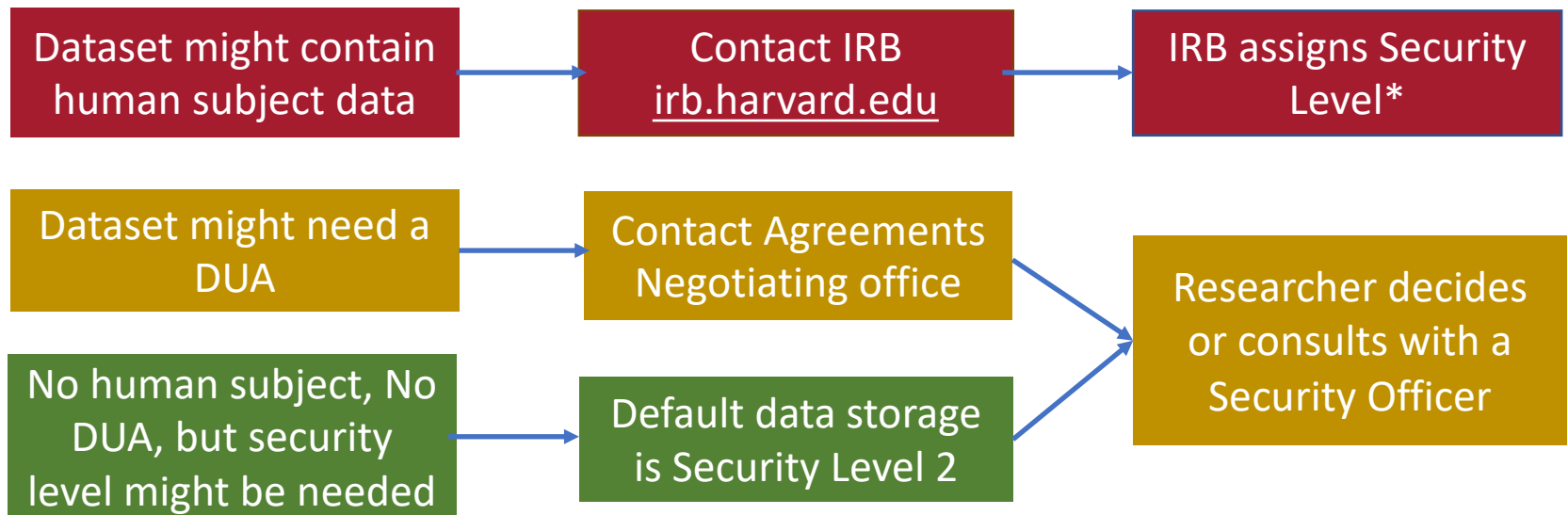| Planning | | Data Acquisition & Collection | | Storage, Security, & Analysis | | Dissemination & Preservation |

Planning for Privacy, Security requirements

- Check whether you need IRB approval
- Check whether you need a Data Use Agreement
- Check whether you need secure storage (Harvard Security Levels, page

Dataset might contain human subject data	→	Contact IRB irb.harvard.edu	→	IRB assigns Security Level*

Dataset might need a DUA	→	Contact Agreements Negotiating office

No human subject, No DUA, but security level might be needed	→	Default data storage is Security Level 2

Researcher decides or consults with a Security Officer

*Escalates to Security Officer if needed

In doubt? Always get written approval

"**Only Written Approval Protects You**"

"… have you ever obtained data from an outside organization, signed an agreement governing how you will use it, and agreed to terms that put you at legal or financial risk if there's a problem? If so, big mistake. Your university employs lawyers who will read and negotiate data use agreements for you and will sign them on behalf of the university if you ask. If you do, this means that your university rather than you personally will bear the legal or financial risk if something goes wrong."

"… with respect to how you treat human subjects, this normally means a simple letter from the IRB or other responsible official. When? Always. For all your research even if you think it is obvious that your research is in compliance with all rules. "

Gary King and Melissa Sands. Working Paper. "How Human Subjects Research Rules Mislead You and Your University, and What to Do About it".

Data Collection, Acquisition, and Use Agreements

Planning		Data Acquisition & Collection		Storage, Security, & Analysis		Dissemination & Preservation

Data Protection Regulations and Policies

- HIPAA (18+ identifiers – alone or in combination datasets) Informed Consent

- FERPA (education information and special protections)

- MA data protection law (security requirements to handle private data from state residents)

- Stem Cell data and Genomics data must be published in approved repository, but also must be de-identified.

- GDPR (General Data Protection Regulation In Europe)

At Harvard:

- Data retention (7 years)

- Data security (5 security levels)

| Planning | | Data Acquisition & Collection | | Storage, Security, & Analysis | | Dissemination & Preservation |

Human Subjects and Animal Research

- Two IRB offices at Harvard, for human subject research:
  - Harvard University Area IRB for main campus and Allston: Committee on the Use of Human Subjects (CUHS)
  - Longwood Area IRB for Medical School, Dental School, and T.H. Chan School of Public Health: Office of Human Research Administration (OHRA)
- For research with animals:
  - HMS Institutional Animal Care and Use Committee
  - FAS Institutional Animal Care and Use Committee

| Planning | | Data Acquisition & Collection | | Storage, Security, & Analysis | | Dissemination & Preservation |

Data Use Agreements

A contract established between the organization providing the data ("provider") and the organization receiving the data ("recipient") that outlines the terms and conditions when sharing confidential, proprietary, or otherwise sensitive data.

Harvard has new DUA guidelines for when:

- You want to bring data from a third party to Harvard for your research

- You want to share data with a third party that you have collected while doing research at Harvard

https://researchdatamanagement.harvard.edu/data-use-agreements

| Planning | → | Data Acquisition & Collection | → | Storage, Security, & Analysis | → | Dissemination & Preservation |

Harvard DUA Guidelines

DUAs must be reviewed and signed by Harvard's Negotiating Offices (OSP, HSPH ORA, HMS ORA):

1. Submit request to the provider to obtain the dataset

2. Provider either drafts a DUA or asks Harvard to draft a DUA

3. Submit request to the appropriate Negotiating Office via Agreements system:
   - https://dua.Harvard.edu/agreeements (Login with HarvardKey)
   - Enter draft and answer questions
   - DUA is reviewed by the Negotiating Office

| Planning | | Data Acquisition & Collection | | Storage, Security, & Analysis | | Dissemination & Preservation |

Subscription Data

The Harvard Library has extensive sets of data that might be useful for your research:

- Consult with an expert librarian via the Harvard Library's Research Help service.

- Search in Hollis or the Harvard Subscription Data Dataverse

- Find text and data mining resource at: https://wiki.harvard.edu/confluence/display/LibraryTech/TDM+@+Harvard

Planning		Data Acquisition & Collection		Storage, Security, & Analysis		Dissemination & Preservation

Storage, Computation, Analysis

Planning		Data Acquisition & Collection		Storage, Security, & Analysis		Dissemination & Preservation

Harvard Security Levels

- **Level 1:** No sensitive data; open data

- **Level 2:** Confidential information by University standards; no material harm; research data without identifiable information

- **Level 3:** Confidential information that could cause material harm (non-level 4 FERPA)

- **Level 4:** High-risk confidential information (SSN, identifiable HIPAA)

- **Level 5:** Information that would cause severe harm (outside the network)

More information:

- [Harvard Research Data Security Policy](#)

- [Harvard Information Security](#)

| Planning | → | Data Acquisition & Collection | → | Storage, Security, & Analysis | → | Dissemination & Preservation |

Storage and Research Computing Options

Four main resources for storage and computing (Level 3, with Level 4 on-demand):

- FAS RC Odyssey: FAS, SEAS, HSPH (limited for other Schools)

- HMS O2: HMS and HSPH

- IQSS RCE: Social science across all Schools

- HBS RCS: HBS only

Other options: (Level 3); Login with HarvardKey

- Harvard's Google Drive

- Harvard's MS OneDrive, SharePoint, Azure (HRCI SharePoint is Level 4)

- Harvard's DropBox

- Harvard's AWS

Planning		Data Acquisition & Collection		Storage, Security, & Analysis		Dissemination & Preservation

Electronic Lab Notebooks (ELNs) and collaborative research tools

- An ELN enables you to track, store, and document all your research:
  - Design
  - Experiments
  - Procedures
  - Data sources and outputs

- Examples: RSpace, LabArchive, LabGuru, Open Science Framework (OSF), Evernote

- Comparison matrix by HMS research data management group: https://datamanagement.hms.harvard.edu/electronic-lab-notebooks

- Harvard is currently evaluating support for general-purpose ELNs useful to all researchers: RSpace and OSF

| Planning | | Data Acquisition & Collection | | Storage, Security, & Analysis | | Dissemination & Preservation |

Computational Notebooks and Code, Workflow tools

- Computational Notebooks:
  - Help document the code and data used in an analysis
  - Jupyter Notebooks: support for most languages
  - RMarkdown and RMarkdown Notebooks for R
- Workflow/Pipeline Tools.
  - Help document and track process order of a entire complex pipeline
  - Consider Drake (R), doit or py-Make (Python); Galaxy or Taverna in life science
  - See https://github.com/pditommaso/awesome-pipeline for a full list
- GitHub: for code repository and version control
- Harvard provides support for most of these tools (e.g., check lists of supported tools at FAS RC VDI OnDemand or IQSS RCE)

Planning		Data Acquisition & Collection		Storage, Security, & Analysis		Dissemination & Preservation

Survey Tools and Data Capture

Harvard supports to following survey tools:

- Qualtrics:
  - Login with HarvardKey
  - Harvard.Qualtrics.com
  - Create, conduct, and analyze surveys
  - Up to Security Level 3

- RedCap:
  - Mostly for Medical Schools and School of Public Health
  - Data capture for (medical) research studies
  - Conduct surveys
  - Up to Security Level 4

Planning → Data Acquisition & Collection → Storage, Security, & Analysis → Dissemination & Preservation

Image Tools; Data Visualization

Harvard Library supports:

- Mirador:
    - https://iiif.harvard.edu/mirador-viewer/
    - Image viewer
    - Supports image annotation and image comparison

- Data visualization:
    - https://library.harvard.edu/services-tools/visualization-support

Planning		Data Acquisition & Collection		Storage, Security, & Analysis		Dissemination & Preservation

Statistics, Data Science, Visualization Consultation and Training

- [FAS Informatics and Research Computing](): Computing support

- [HMS Research Computing services](): Computing support

- [Chan Bioinformatics Core](): Biostatistics support

- [HBS Research Computing & Training](): Computing and statistics

- [IQSS Data Science Services](): Research design, stats, machine learning

- [Digital Scholarship and Support Group](): Digital humanities support

- [Wolbach Library at the Center for Astrophysics](): Astronomy and astrophysics support

- [Center for Geographic Analysis](): Geospatial research support

- [Program on Survey Research](): Survey creation and analysis support

| Planning | → | Data Acquisition & Collection | → | Storage, Security, & Analysis | → | Dissemination & Preservation |

Data Sharing and Preservation

| Planning | → | Data Acquisition & Collection | → | Storage, Security, & Analysis | → | Dissemination & Preservation |

Recent Increase in Data Sharing Awareness among Researchers

Introducing the PLOS Open Data Collection

Credit: KamiPhuc
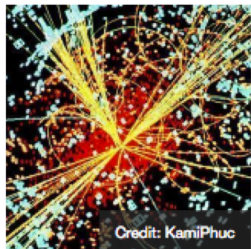
Introducing the PLOS Open Data Collection
Daniella Lowenberg, Amy Ross, Emma Ganley

PLOS Collections Blog: 10 Nov 2016

Data sharing in a modern world
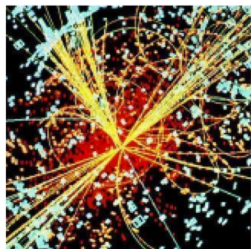
Data sharing in a modern world; well, maybe not so modern
Melissa Haendel & Nicole Vasilevsky

PLOS Collections Blog: 10 Nov 2016

Open Data Articles

Ethical Challenges of Big Data in Public Health
Effy Vayena, Marcel Salathé, Lawrence C. Madoff, John S. Brownstein

PLOS Computational Biology: 09 Feb 2015

Ten Simple Rules for the Care and Feeding of Scientific Data
Alyssa Goodman, Alexander W. Blocker, Christine L. Borgman, Kyle Cranmer, Merce Crosas, Rosanne Di Stefa…

PLOS Computational Biology: 24 Apr 2014
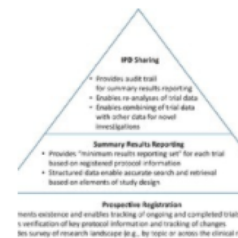
Sharing Individual Participant Data (IPD) within the Context of the Trial Reporting System (TRS)
Deborah A. Zarin, Tony Tse

PLOS Medicine: 19 Jan 2016

http://collections.plos.org/open-data

| Planning | → | Data Acquisition & Collection | → | Storage, Security, & Analysis | → | Dissemination & Preservation |

Social Science Journals Data Policies

Percentage of Journals by Strictness of Data Policy
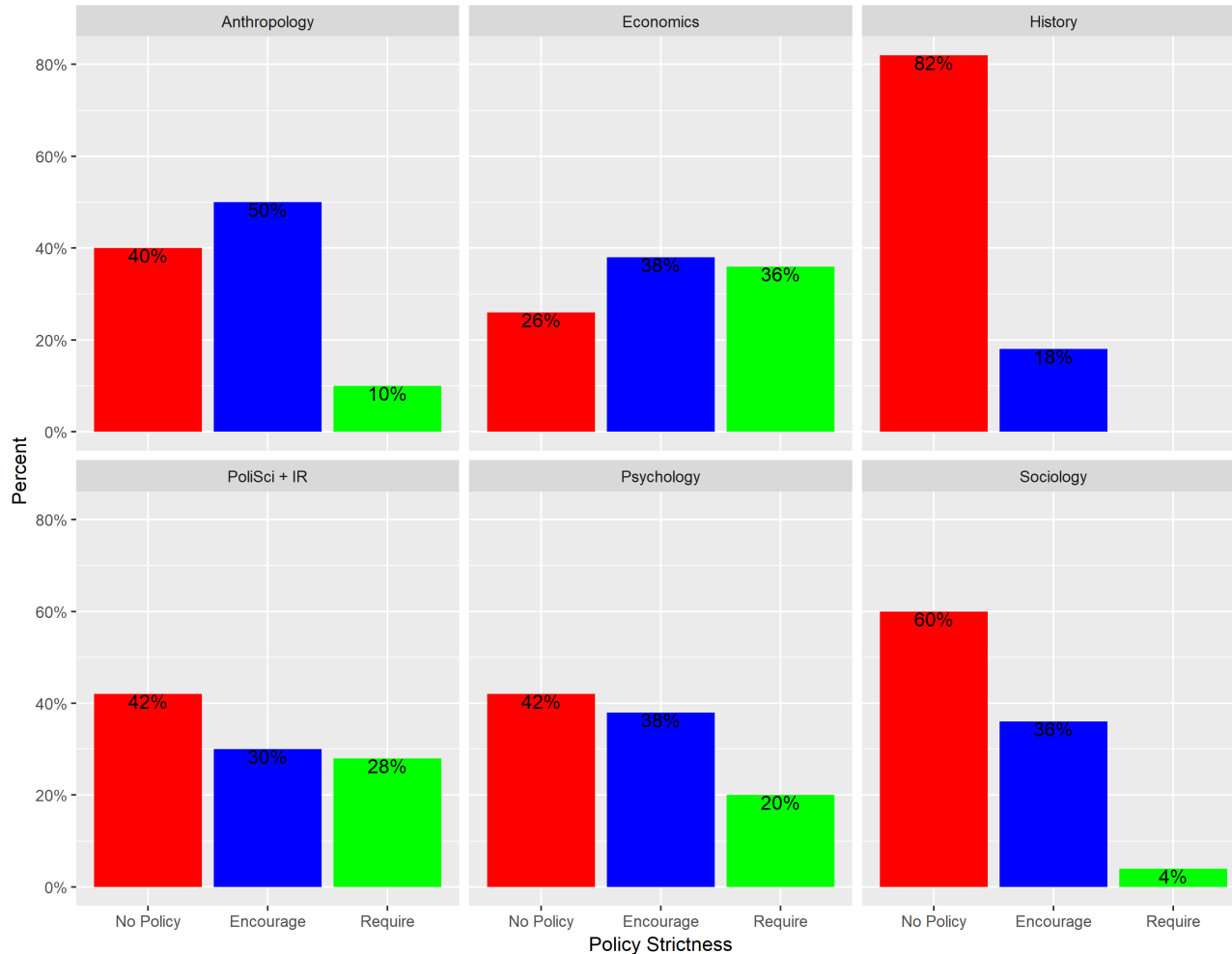
Crosas et al 2018, *preprint*, Data Policies of highly-ranked social science journals

Planning → Data Acquisition & Collection → Storage, Security, & Analysis → Dissemination & Preservation

When Sharing Data ...

- Make open research data the default: **"**as open as possible, as closed as needed"

- Comply with funding organizations and journals requirements

- Get credit for your data through data citation

- Reward researchers when using their data by giving them credit

- Use trusted data repositories aligned with FAIR data principles — Findable, Accessible, Interoperable, and Reusable

Planning		Data Acquisition & Collection		Storage, Security, & Analysis		Dissemination & Preservation

Harvard Dataverse repository:
https://dataverse.harvard.edu

HARVARD
Dataverse

Deposit and share your data. Get academic credit.

Harvard Dataverse is a repository for research data. Deposit data and code here.

90,252 datasets   **8,161,289** downloads

Add a dataset +

Organize datasets and gather metrics in your own repository.

A dataverse is a container for all your datasets, files, and metadata.

3,337 dataverses

Add a dataverse +

Find data across research fields, preview metadata, and download files

| Search over 90,200 datasets... | 🔍 Find |

Browse by subject

Agricultural Sciences 1,264	Computer and Information Science 931	Medicine, Health and Life Sciences 3,011
Arts and Humanities 804	Earth and Environmental Sciences 1,480	Physics 516
Astronomy and Astrophysics 516	Engineering 416	Social Sciences 38,674
Business and Management 428	Law 276	
Chemistry 184	Mathematical Sciences 209	

Harvard Dataverse Repository

- Generates a data citation which you can use in the reference section of your scholarly article

- Provides support for standard metadata, plus custom metadata

- Enables tiered access to data:
  - Fully Open, CC0
  - Register to access; Guestbook
  - Restricted with DUA

- Allows multiple versions of a dataset

- Aligns with FAIR data principles and Joint Declaration of Data Citation Principles

| Planning | Data Acquisition & Collection | Storage, Security, & Analysis | Dissemination & Preservation |

HARVARD
Dataverse

Replication Data for: Cost of Compliance, Autocratic Time Horizon, and Investment Treaty Formation `Version 1.0`

Description	This paper investigates the domestic political factors that shape the participation of autocratic regimes in Bilateral Investment Treaties (BITs). We argue that autocratic time horizon positively affects governments' motives to sign BITs by influencing the costs of complying with investor protection standards included in the treaties. These treaty provisions severely constrain discretionary policy maneuvers that are critical to autocratic survival. Autocratic regimes expecting to rule for a considerable time period are willing to relinquish some discretionary policy space in the interest of enhancing the credibility of their investor protection commitment – and hence promoting investment inflows. However, autocratic governments with short time horizons rely heavily on discretionary policy maneuvers to stabilize their grip on power, and are likely to infringe on investors' interests to extract resources to ensure their political survival, making the costs of compliance with BITs too high to bear. Using a country-dyad dataset of BIT signatures from 1971 to 2009, we find strong support for our argument.
Subject	Social Sciences
Related Publication	Chen, Jia, and Fangjin Ye. Forthcoming. "Cost of Compliance, Autocratic Time Horizon, and Investment Treaty Formation." Political Research Quarterly.

A dataset can contain any type of research data, documentation, code

| Files | Metadata | Terms | Versions |

Search this dataset...   🔍 Find

1 to 5 of 5 Files

☐   ⬇ Download ▾

☐   **home_full_host.tab**
Tabular Data - 141.9 MB - Jan 13, 2019 - 0 Downloads
130 Variables, 193662 Observations -
UNF:6:lQhb/Bf3hPwddJtt/3TZ9w==
⠿ Explore   ⬇ Download ▾

☐   **pr_isds.tab**
Tabular Data - 90.2 KB - Jan 13, 2019 - 0 Downloads
13 Variables, 1101 Observations -
UNF:6:N/UdJr+gAAPh22f2omjsYQ==
⠿ Explore   ⬇ Download ▾

☐   **replication do file.do**
application/x-stata-syntax - 13.4 KB - Jan 13, 2019 - 0
Downloads
MD5: 2cf24c50ee2eec3981aa6ec91529fb15
⬇ Download

Who uses Harvard Dataverse and what for?

Election Data Archive (Steve Ansolabehere, Government Department)

Ebola Data (Pardis Sabeti, Department of Organismic and Evolutionary Biology)

Robert Sampson's Data (Sociology Department)

Supernova Data (Alicia Soberderg, Astronomy Department, CfA)

New Data Curation Services

Collaborative services offered by Harvard Library and IQSS
Anticipated launch: Late-fall 2019

PUBLISH & SHARE

Organize & share data in repository

Harvard Dataverse data repository & DASH

Data curation services

Consultations, referrals & best practices

HARVARD DATAVERSE REPOSITORY

- FAIR data
- Free data deposits
- Self-curation
- DOIs
- Data citations

CONSULTATION (FALL 2019)

- Free consultation & assessment
- Fee-based extended consultation services

DATA CURATION (FALL 2019)

- Dataverse setup & file ingest
- Ongoing dataverse administration & data curation
- Custom data curation services

Planning → Data Acquisition & Collection → Storage, Security, & Analysis → Dissemination & Preservation

Other Data Repositories

A list of discipline specific repositories can be found at Springer-Nature website or at https://www.re3data.org/

SPRINGER NATURE

< Authors
Research data

Recommended Repositories

Submitting to Research Data Support

Who can use Research Data Support

In general, data should be submitted to discipline-specific, community-recognised repository where possible, or to generalist repositories if no suitable community resource is available.

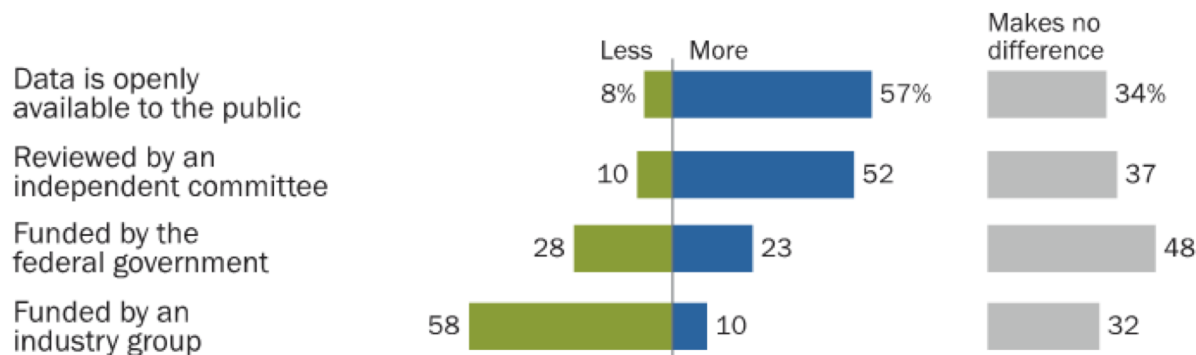| Planning | | Data Acquisition & Collection | | Storage, Security, & Analysis | | Dissemination & Preservation |

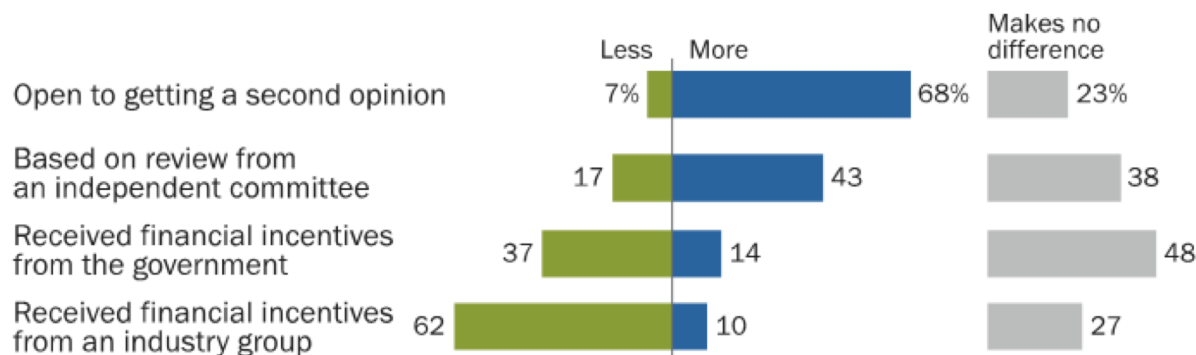**Majority of Americans say they are more apt to trust research when the data is openly available**

% of U.S. adults who say when they hear each of the following, they trust scientific research findings ...

	Less	More	Makes no difference
Data is openly available to the public	8%	57%	34%
Reviewed by an independent committee	10	52	37
Funded by the federal government	28	23	48
Funded by an industry group	58	10	32

% of U.S. adults who say when they hear each of the following, they trust a science practitioner's recommendation ...

	Less	More	Makes no difference
Open to getting a second opinion	7%	68%	23%
Based on review from an independent committee	17	43	38
Received financial incentives from the government	37	14	48
Received financial incentives from an industry group	62	10	27

Note: Respondents who did not give an answer are not shown.
Source: Survey conducted Jan. 7-21, 2019.
"Trust and Mistrust in Americans' Views of Scientific Experts"

PEW RESEARCH CENTER

New Results from Pew Research Center Survey

August 2, 2019

https://www.pewresearch.org/science/2019/08/02/americans-say-open-access-to-data-and-independent-review-inspire-more-trust-in-research-findings/

Thanks! Questions?