

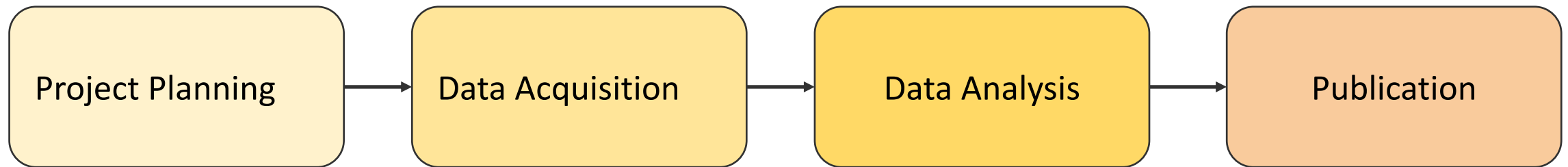
Responsible and FAIR (Findable, Accessible, Interoperable, Reusable) Research Data Management at Harvard University

Mercè Crosas, Ph.D.

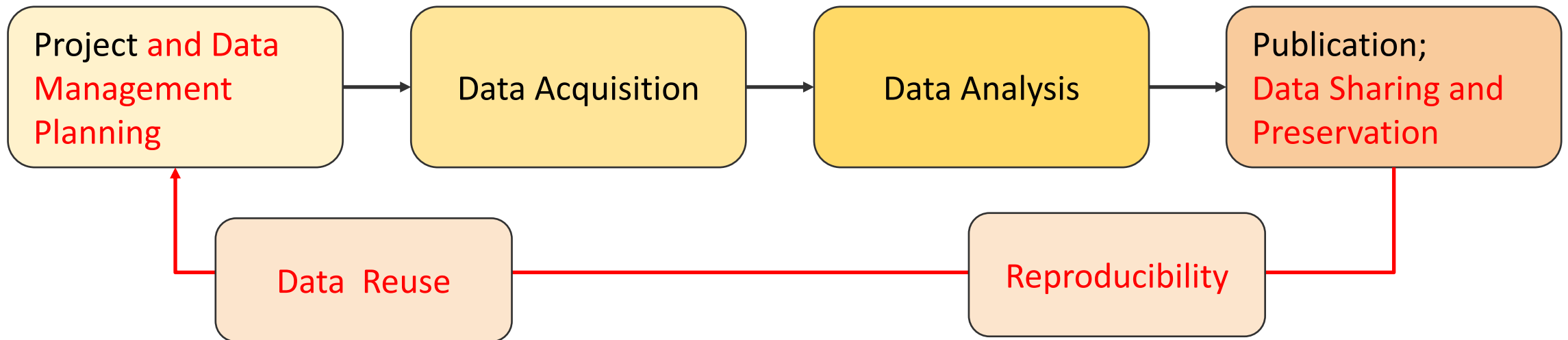
Harvard University's Research Data Officer, Office of Vice Provost for Research
Chief Data Science and Technology Officer, Institute for Quantitative Social Science

@mercecrosas

The Old Data Lifecycle

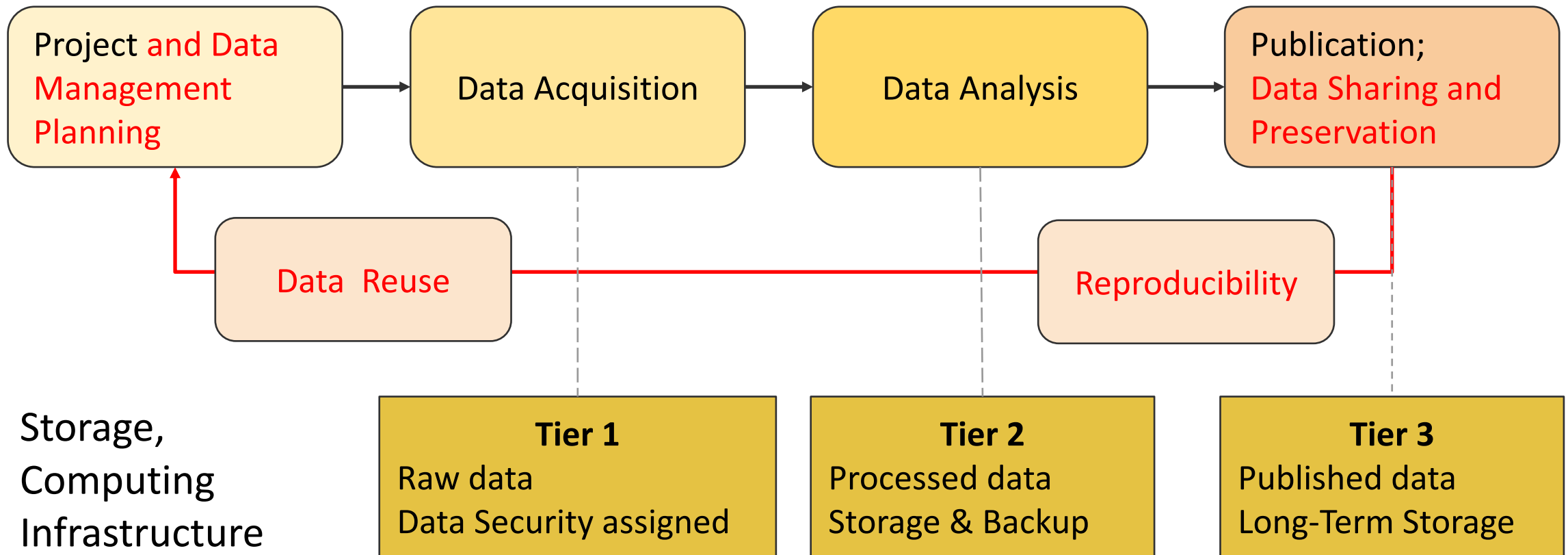


The New Data Lifecycle: *Data as a Product of Research*



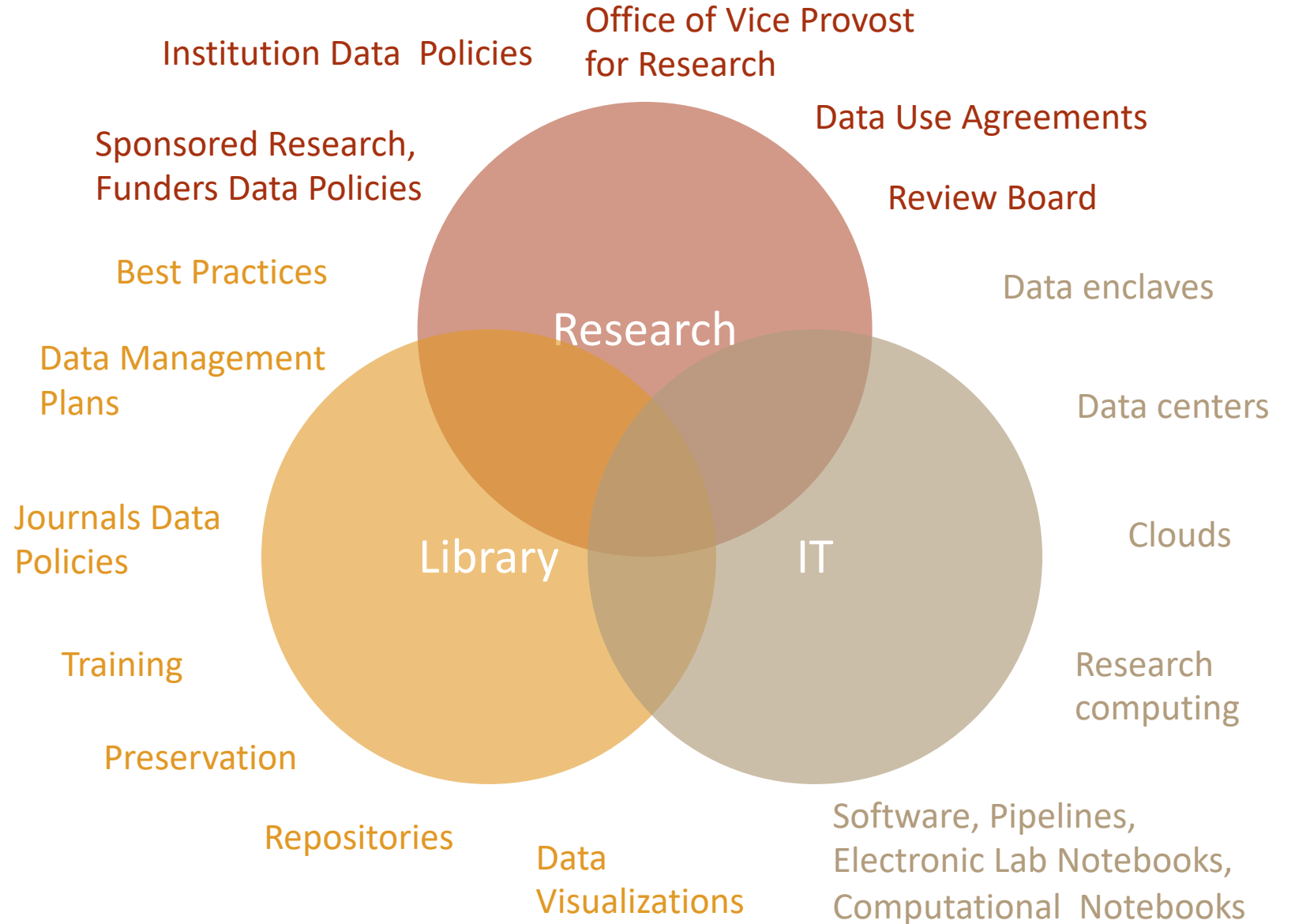
- Funders and Journals Data Sharing Policies
- Credit for your data

The New Data Lifecycle: *Data as a Product of Research*



Collaboration is Key

- Policies, compliance
- Procedures, guidelines
- Best practices, outreach
- Training, support
- Repositories, tools
- Computation and storage infrastructure



A One-stop Resource

<http://researchdatamanagement.harvard.edu>

- Launched in January 2019
- A Harvard-wide resource
- Supports the new data lifecycle
- Links to other relevant Harvard sites
- News and Data Shares blog
- Content provided by groups across Schools, Library, IT, and Office of Vice Provost for Research
- [still growing and improving]

HARVARD UNIVERSITY HARVARD.EDU

Research Data Management @Harvard


Home Vision Data Lifecycle Policies Resources Contacts

A reference guide with information and resources to help you manage your research data

DATA LIFECYCLE


- Planning Data Management**
How can I best manage my data throughout the lifecycle of my research to save time and money in the future?
 - Data Management Plans (DMPs)
 - DMP requirements and tools
 - What research objects should be tracked and documented
- Data Acquisition and Collection**
How can I acquire data in an efficient and ethical way, and how can I ensure that my data is used appropriately?
 - Data Use Agreements (DUAs)
 - Institutional Review Boards (IRB and IACUC)
 - Subscription data
- Storage, Security, and Analysis**
What are my options for effectively organizing, storing, securing, computing, and analyzing my research data?
 - Data security
 - Computing, research methods, data science, and viz support
 - Electronic Lab Notebooks
- Dissemination and Preservation**
Why is it worthwhile to share my data? What do funders and journals require? Can I get help with data curation?
 - Data repositories
 - Open Access
 - Data citation, FAIR principles
 - Data disposal

LATEST NEWS

 **New Data Use Agreement Guidance for Harvard Researchers**
January 4, 2019
As part of a larger initiative to improve transparency of research compliance processes, the University has introduced a tool for submission, review, and management of Data Use Agreements (DUAs). In support of recently [published guidance by the Office for the Vice Provost for Research](#) on the management of DUAs, researchers sharing data can use the system to request a DUA review, correspond with the DUA reviewer, track the status of review, and manage active DUAs (including amendments and extensions...
[Read more](#)

DATA SHARES

One Step Closer to the “Paper of the Future”
March 13, 2019

 *Catherine Zucker is a fourth-year PhD student in astronomy at Harvard. Her work focuses on understanding the structure of our Milky Way Galaxy, through the combination of observations, numerical simulations, statistics, and data visualization. She is advised by Professors Alyssa Goodman and Douglas Finkbeiner. [Alyssa Goodman](#) is the Robert Wheeler Willson Professor of Applied Astronomy,*

Working Groups to Foster Collaboration and Diverse Input

RDM Support WG

- **Creates and coordinates website, use cases, DMP templates, services, best practices**
- Chairs: Ceilyn Boyd, Julie Goldman
- 8 members from HL, HBS RCS, Chan Bioinformatics Core, HMS RDM, IQSS, OVPR

Electronic Lab Notebooks WG

- **Coordinates ELN rollout, support and training**
- Chairs: Mercè Crosas, Alan Wolf, Mason Miranda
- 6 members from OVPR, HUIT, HMS RITS, HL

Data Use Agreements Operations WG

- **Develops and implements DUAs policies and procedures incoming and outgoing data**
- Chairs: Jennifer Ponting, Alisa Jahns
- 16 members from OVPR, OSP, HUIT, HMS-ORA, SPH-SPA, SEAS, FAS, GSE, HMS Health Care Policy

Data Safety and Security WG

- **Develops a University-wide tracking and compliance system for data security**
- Chairs: Alisa Jahns, Rachel Talentino
- 9 members from OVPR, HUIT, HMS, GSE, SPH, HMS Health Care Policy

Open source research data repository software



A software, a community, many repositories

The Dataverse Software

<http://dataverse.org>

- Developed since **2006** at Harvard's Institute for Quantitative Social Science
- **89** contributors, most external to Harvard
- **> 1000** pull requests in GitHub
- **12** releases a year
- **43** installations around the world

The Harvard Dataverse Repository

<http://dataverse.harvard.edu>

- Open to all researchers, all disciplines
- **30,000** datasets deposited
- **+ 50,000** datasets harvested from other Repositories
- **250** new datasets added per month
- **7 million** file downloads

A Rich Set of Features

- **Data citation:** credit as an incentive to share data
- **Metadata:** find and reuse data
 - Data Documentation Initiative (DDI)
 - DataCite (+ OpenAire)
 - Dublin Core
 - Schema.org
- **Versioning** for dataset and files
- **Tiered access to data:** guestbook, terms of use and licenses, file restrictions
- **Integration** with data exploration tools
- **Customization** and branding of your own dataverse (your collection of datasets)
- Extensive **APIs**
- OAI-PMH for metadata **harvesting** from one repository to another


Three Years Ago ...

nature > scientific data > comment > article

SCIENTIFIC DATA 

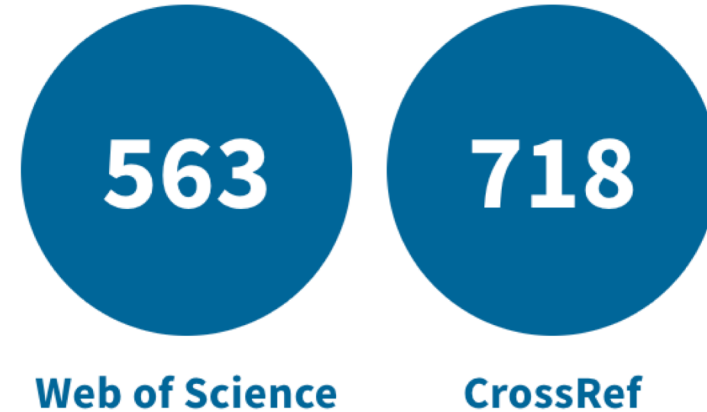
Comment | [OPEN](#) | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons  - [Show fewer authors](#)

Scientific Data **3**, Article number: 160018 (2016) | [Download Citation](#) ↓






Total citations



Online attention



Altmetric score (what's this?)

-  Tweeted by **1177**
-  Blogged by **69**
-  On **16** Facebook pages
-  Mentioned in **6** Google+ posts
-  Picked up by **81** news outlets

 [Show more](#)

”The FAIR Principles put specific emphasis on enhancing the ability of **machines to automatically find and use the data**, in addition to supporting its **reuse by individuals**. ”

Wilkinson et al. 2016 “The FAIR Guiding Principles for scientific data management and stewardship” Scientific Data

The FAIR Guiding Principles

- **To be Findable:**

- F1. (meta)data are assigned a **globally unique and persistent identifier**
- F2. data are described with rich **metadata** (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are **registered or indexed** in a searchable resource

- **To be Accessible:**

- A1. (meta)data are retrievable by their identifier using a **standardized communications protocol**
- A1.1 the protocol is open, free, and universally implementable
- A1.2 the protocol allows for **an authentication and authorization** procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

- **To be Interoperable:**

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use **vocabularies** that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

- **To be Reusable:**

- R1. meta(data) are richly described with a plurality of accurate and relevant **attributes**
- R1.1. (meta)data are released with a clear and accessible data **usage license**
- R1.2. (meta)data are associated with detailed **provenance**
- R1.3. (meta)data meet **domain-relevant community standards**

Each dataset has a landing page with human-readable metadata and a global persistent identifier

Harvard Dataverse

AMERICAN JOURNAL of POLITICAL SCIENCE

American Journal of Political Science (AJPS) Dataverse (Midwest Political Science Association) ajps.org

Harvard Dataverse > American Journal of Political Science (AJPS) Dataverse >

Replication Data for: Strategic Spending: Does Politics Influence Election Administration Expenditure?

Version 1.1 **Blue**

Pope, JoEllen; Kropf, Martha; Shepherd, Mary Jo; Mohr, Zachary, 2019, "Replication Data for: Strategic Spending: Does Politics Influence Election Administration Expenditure?", <https://doi.org/10.7910/DVN/8KNF31>, Harvard Dataverse, V1, UNF:6:Axve/NFbJVvU/qcPe+IRA= [fileUNF]

Cite Dataset [Learn about Data Citation Standards](#)

Description Abstract: Recently, election administration has been an important part of the national and global conversation about the results of elections. The important issue of election administration spending has not been examined extensively, and the influence of politics on election administration spending [+ More](#)

Subject Social Sciences

Keyword Elections, Administration, Budget, County

Related Publication Mohr, Zachary, JoEllen Pope, Martha Kropf, and Mary Jo Shepherd. [date]. "Strategic Spending: Does Politics Influence Election Administration Expenditure?" American Journal of Political Science. Forthcoming. <http://ajps.org/>

Data Quality [+ More](#)

Notes This dataset underwent an independent verification process that replicated the tables and figures in the primary article. For the supplementary materials, verification was performed solely for the successful execution of code. The verification process was carried out by the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill.

Files Metadata Terms Provenance Versions

View: Search this dataset...

1 to 4 of 4 Files

- AJPS_strategic_spending_original.tab** **Blue** **Data**
/example/directory/structure/
Tabular Data - 163.7 KB - Feb 12, 2019 - 2 Downloads
30 Variables, 1100 Observations - UNF:6:Axve...IRA= [+](#)
Data file for Strategic Spending
- CODEBOOK.pdf** **Blue** **Documentation**
/example/directory/structure/
Adobe PDF - 432.1 KB - Feb 12, 2019 - 5 Downloads
MD5:082...463 [+](#)
- readme.txt** **Blue**
/example/directory/structure/
Plain Text - 315 B - Feb 12, 2019 - 6 Downloads
MD5:947...d36 [+](#)
Description of file contents
- z_mohr_replication_do.do** **Blue**
/example/directory/structure/
application/x-stata-syntax - 12.8 KB - Feb 12, 2019 - 3 Downloads
MD5:1ea...f28 [+](#)

Mercè Crosas, IQSS, OVPR, Harvard University @mercecrosas

DataCite DOIs in data citation (**Findable** and **Accessible**)

Make Data Count (**Reusable** metric)

Citation and discoverable metadata using DataCite, schema.org, Dublin Core, DDI standards (**Findable**, **Accessible** and **Reusable**)

More metadata, including domain-specific (**Reusable**)

Terms with license or Data Use Agreement (**Reusable**)

PROV metadata (**Reusable**)

Each dataset has machine-readable metadata

```
<title>Replication Data for: Strategic Spending: Does Politics Influence Election Administration Expenditure? - American Journal of Political Science (AJPS) Dataverse</title>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<meta name="DC.identifier" content="doi:10.7910/DVN/8KNF3I" />
<meta name="DC.type" content="Dataset" />
<meta name="DC.title" content="Replication Data for: Strategic Spending: Does Politics Influence Election Administration Expenditure?" />
<meta name="DC.date" content="2019-03-11" />
<meta name="DC.publisher" content="Harvard Dataverse" />
<meta name="DC.description" content="Abstract: Recently, election administration has been an important part of the national and global conversation about the results of elections. The important issue of election administration spending has not been examined extensively, and the influence of politics on election administration spending levels has not been examined in the United States. While theories of voter turnout and policy preference suggest that politics should influence election administration spending levels in the counties that administer elections, to our knowledge, there has been no evidence produced to support a partisan election administration expenditure effect. This research finds that Republican county commissions in North Carolina spend significantly less on election administration once the county electorate is a sufficient Republican majority. The paper presents a novel model and method for estimating election administration spending and calls for additional research to examine the outcomes of these significant differences in spending on election administration." />
<meta name="DC.creator" content="Pope, JoEllen" />
<meta name="DC.creator" content="Kropf, Martha" />
<meta name="DC.creator" content="Shepherd, Mary Jo" />
<meta name="DC.creator" content="Mohr, Zachary" />
<meta name="DC.subject" content="Social Sciences" />
<script type="application/ld+json">
{ "@context": "http://schema.org", "@type": "Dataset", "@id": "https://doi.org/10.7910/DVN/8KNF3I", "identifier": "https://doi.org/10.7910/DVN/8KNF3I", "name": "Replication Data for: Strategic Spending: Does Politics Influence Election Administration Expenditure?", "creator": [{"name": "Pope, JoEllen", "affiliation": "University of North Carolina at Charlotte"}, {"name": "Kropf, Martha", "affiliation": "University of North Carolina at Charlotte"}, {"name": "Shepherd, Mary Jo", "affiliation": "University of North Carolina at Charlotte"}, {"name": "Mohr, Zachary", "affiliation": "University of North Carolina at Charlotte"}], "author": [{"name": "Pope, JoEllen", "affiliation": "University of North Carolina at Charlotte"}, {"name": "Kropf, Martha", "affiliation": "University of North Carolina at Charlotte"}, {"name": "Shepherd, Mary Jo", "affiliation": "University of North Carolina at Charlotte"}, {"name": "Mohr, Zachary", "affiliation": "University of North Carolina at Charlotte"}], "datePublished": "2019-02-12", "dateModified": "2019-03-11", "version": "1", "description": ["Abstract: Recently, election administration has been an important part of the national and global conversation about the results of elections. The important issue of election administration spending has not been examined extensively, and the influence of politics on election administration spending levels has not been examined in the United States. While theories of voter turnout and policy preference suggest that politics should influence election administration spending levels in the counties that administer elections, to our knowledge, there has been no evidence produced to support a partisan election administration expenditure effect. This research finds that Republican county commissions in North Carolina spend significantly less on election administration once the county electorate is a sufficient Republican majority. The paper presents a novel model and method for estimating election administration spending and calls for additional research to examine the outcomes of these significant differences in spending on election administration."], "keywords": ["Social
```

Dublin Core meta-tags for citation metadata
(Findable and Accessible)

Schema.org JSON-LD
(Findable in Google Dataset Search)

HARVARD
Dataverse

Search ▾ About User Guide Support Sign Up Log In

Files **Metadata** Terms Versions

Search this dataset...

Restricted data files (Authentication and Authorization needed)

<input type="checkbox"/>			<input type="button" value="Request Access"/>
<input type="checkbox"/>		<p>cameo_candidate_master_harvester_pairs.csv Plain Text - 634.0 KB - Jun 21, 2015 - 0 Downloads MD5: 8f52e231fb316004c9668e65a6c7aa02 Please see "description_of_cameo_candidate_master_harvester_pair.txt"</p> <p>CAMEO candidate master harvester pairs Data</p>	<input type="button" value="Request Access"/>
<input type="checkbox"/>		<p>cameo_course_listings.csv Plain Text - 8.0 KB - Jun 21, 2015 - 0 Downloads MD5: 653639926109cc2dd0021db1cd7a3f14 Please see "description_of_cameo_course_listings.txt"</p> <p>CAMEO course listings Data</p>	<input type="button" value="Request Access"/>

A dataset can contain open and/or restricted data files; Metadata is always accessible

Landing Page with metadata, DUA, license **(Accessible)**

Open data files (direct download)

<input type="checkbox"/>			<input type="button" value="Download"/>
<input checked="" type="checkbox"/>		<p>AJPS_strategic_spending_original.tab Tabular Data - 163.7 KB - Feb 12, 2019 - 3 Downloads 30 Variables, 1100 Observations - UNF:6:Avxe/NFbJVU/lqcPe+IRA== Data file for Strategic Spending</p>	<input type="button" value="Explore"/> <input type="button" value="Download"/>
<input type="checkbox"/>		<p>CODEBOOK.pdf Adobe PDF - 432.1 KB - Feb 12, 2019 - 5 Downloads MD5: 082283da72122f19f23d43dc0a040463</p>	
<input type="checkbox"/>		<p>readme.txt Plain Text - 315 B - Feb 12, 2019 - 6 Downloads MD5: 9472652d0a7d02ae1c0095ab11bfd36 Description of file contents</p>	
<input type="checkbox"/>		<p>z_mohr_replication_do.do application/x-stata-syntax - 12.8 KB - Feb 12, 2019 - 3 Downloads MD5: 1ea8fbe91cf2c752c4db8645f4ff4f28</p>	<input type="button" value="Download"/>

Download data and metadata using domain-relevant standards **(Reusable)**

Replication Data for: Strategic Spending: Does Politics Influence Election Administration Expenditure?

AJPS_strategic_spending_original.tab

Pope, JoEllen; Kropf, Martha; Shepherd, Mary Jo; Mohr, Zachary, 2019, "Replication Data for: Strategic Spending: Does Politics Influence Election Administration Expenditure?", <https://doi.org/10.7910/DVN/8KNF3I>, Harvard Dataverse, V1, UNF:6:Avxe/NFbJVvU/IqcPe+IRA== [fileUNF]

Q 30 Results Download

Chart View Table View

ID	Name	Label
19184087	counties	County Names string
19184099	county	County Names
19184102	cpi	Consumer Price Index Research Series Using Current Methods (CPI-U-RS)1977=100

Records Per Page 10

Variable county: County Names

Values	Categories	N
48	Hyde	
94	Washington	
6	Avery	
67	Onslow	
12	Halifax	

ID	Name	Label
19184098	totvapunder25	total voting age population under 25
19184090	vap	voting age poplation
19184103	vemodel	voting equipment model name string

Records Per Page 10

Summary Statistics

Cases	N
	1100
	0
Maximum	763682
Minimum	2838
	65065.06818181813

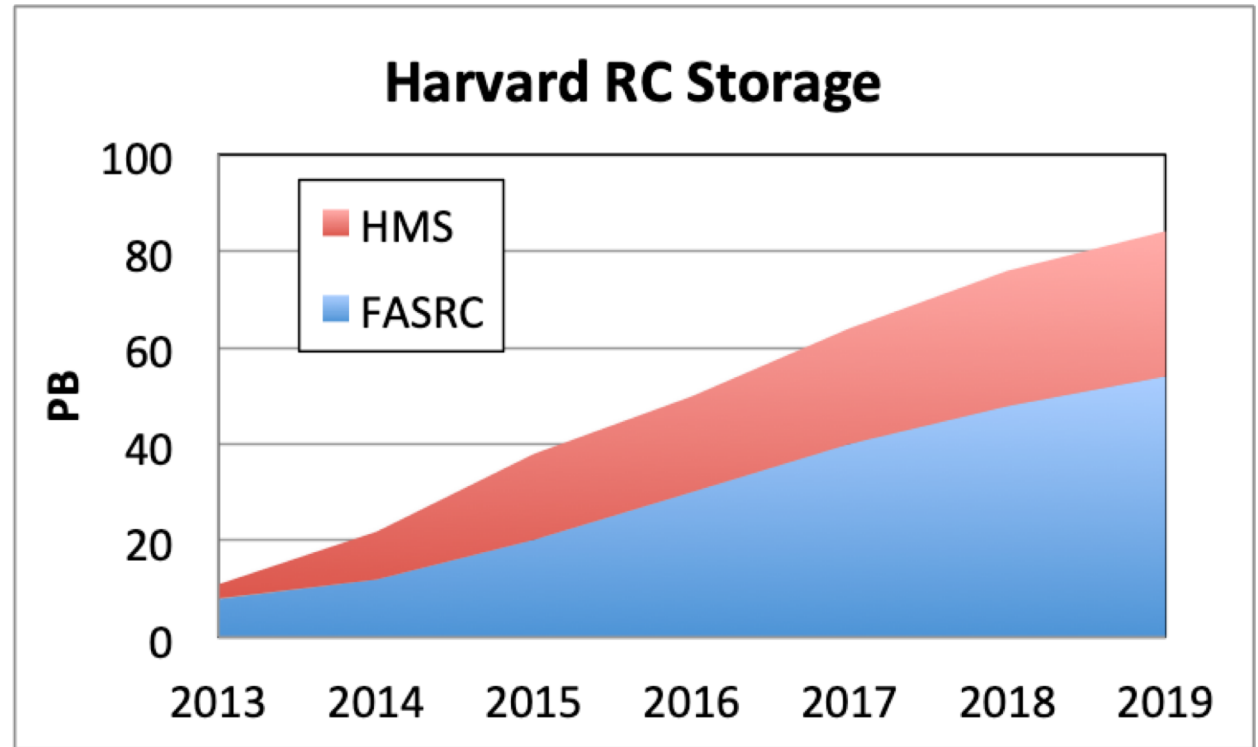
Human and machine readable metadata at the variable level

Machine-readable Variable description from DDI (Interoperable and Reusable)

Summary Statistics in DDI, automatically calculated upon data upload (Interoperable and Reusable)

Next Challenges

- The continued growth of research data
- Complex privacy and security requirements
- Data quality: new curation services (Library, IQSS)

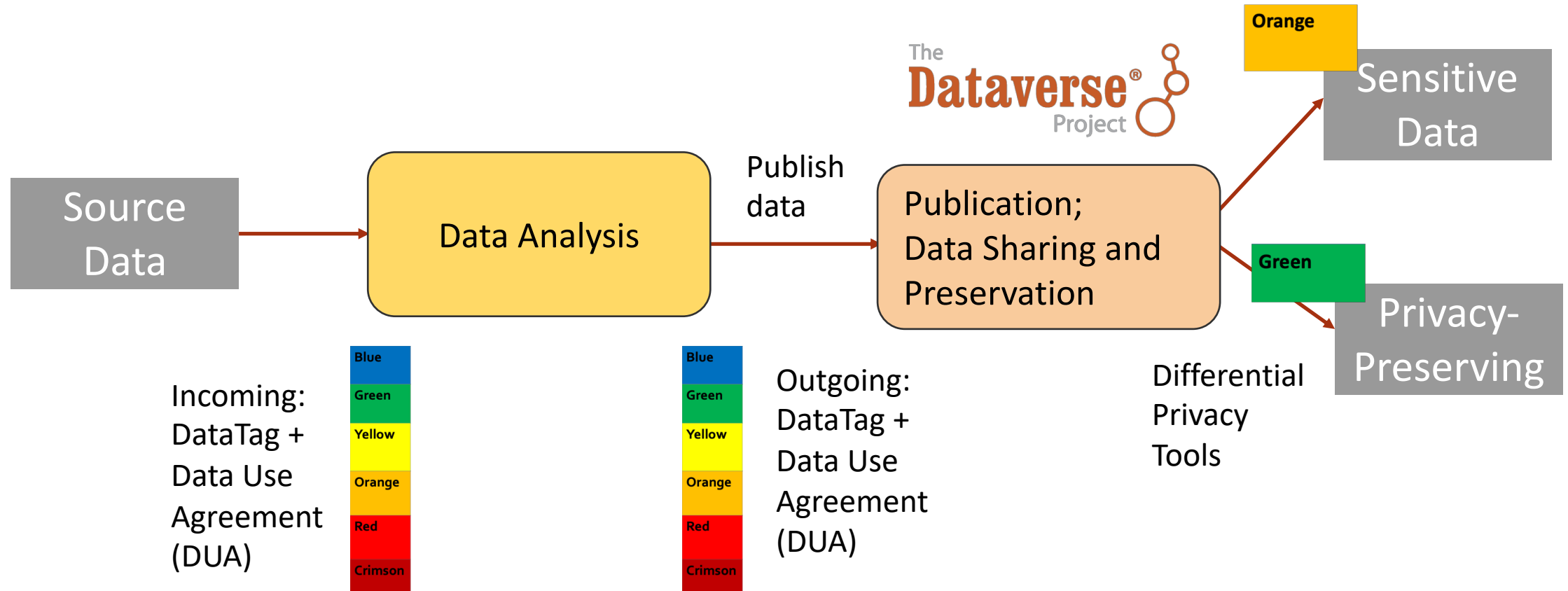


DataTags Facilitate Sharing Sensitive Data Responsibly

Blue	Public			
Green	Public Accountable	Register		
Yellow	Restricted Not Sensitive	Approval Needed	Click-thru Data Use Agreement (DUA)	Encrypted transmit
Orange	Restricted Sensitive	Approval Needed	Signed DUA	Encrypted transmit Encrypted storage
Red	Restricted High Sensitive	Approval Needed	Signed DUA Two-factor Auth	Encrypted transmit Encrypted storage
Crimson	Restricted Max Sensitive	Approval Needed	Signed DUA Two-factor Auth	Encrypted transmit Multi-encrypted storage

Sweeney, Crosas, Bar-Sinai, 2015. *Sharing Sensitive Data with Confidence: The DataTags System*, Technology Science

Integrating DataTags with Dataverse



The Harvard Data Commons: A Proposed Integrated System



Check-in unpublished dataset in repository

Persistent ID
Metadata
DUA + Security + Access
Provenance + code/container

Publish dataset

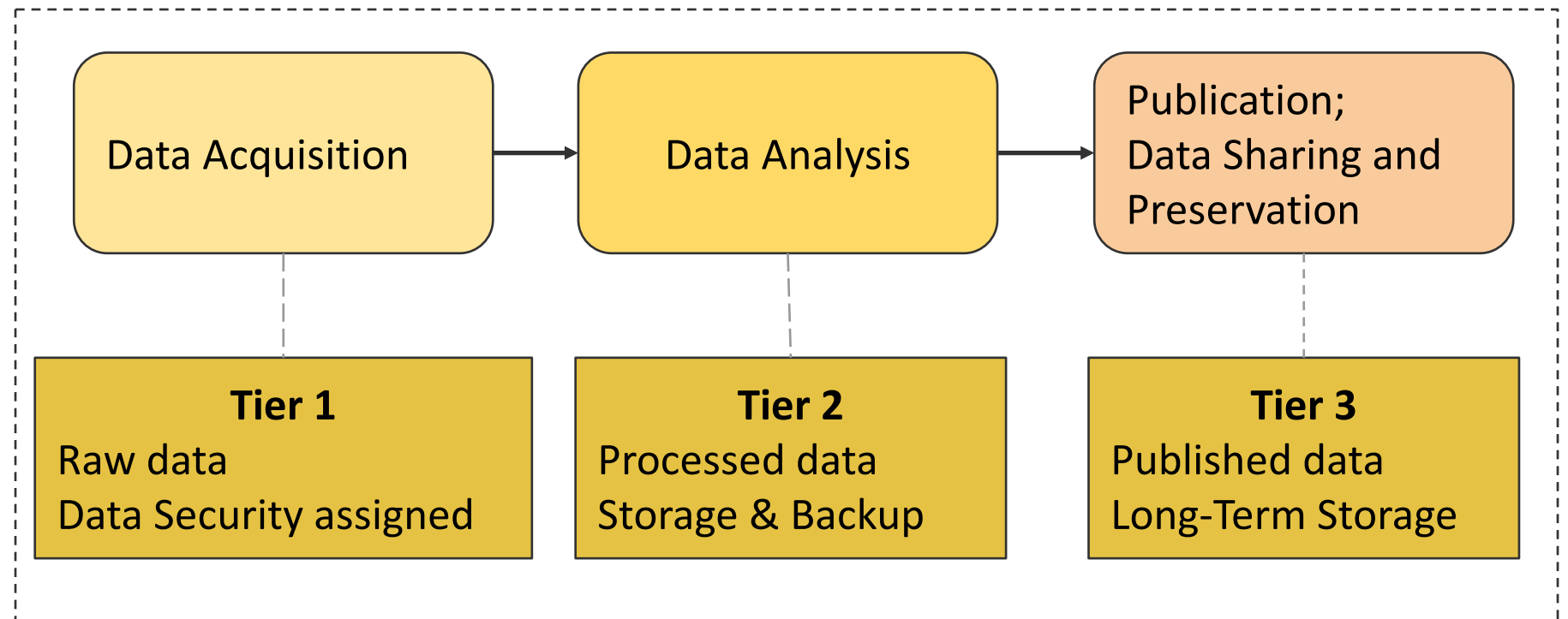
Electronic Lab Notebooks



Big data transfer



Policy-Based, Scalable File and Object System



Thanks!

Mercè Crosas, Ph.D.

Harvard University's Research Data Officer, Office of Vice Provost for Research
Chief Data Science and Technology Officer, Institute for Quantitative Social Science
[@mercecrosas](https://twitter.com/mercecrosas)