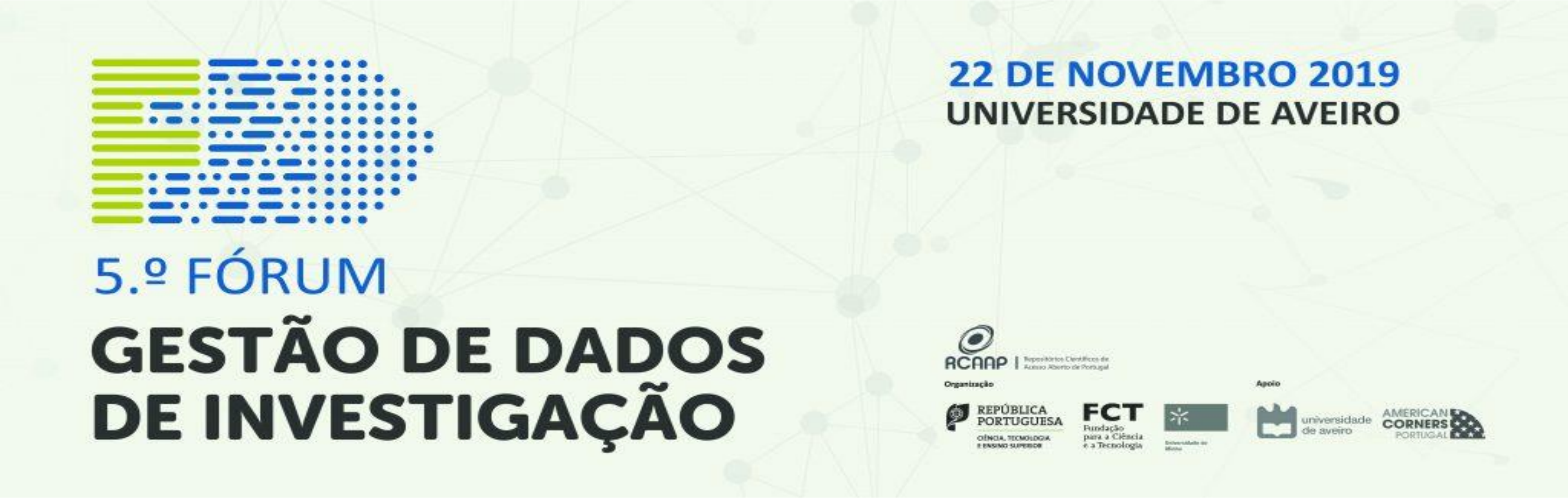


22 DE NOVEMBRO 2019
UNIVERSIDADE DE AVEIRO

5.º FÓRUM

GESTÃO DE DADOS DE INVESTIGAÇÃO

Organização
Apoio

Research Data Curation, Management, Sharing and Archiving

Jonathan Crabtree

Director of Cyberinfrastructure

Odum Institute for Research in Social Science

THE H. W. ODUM INSTITUTE FOR RESEARCH IN SOCIAL SCIENCE

- Founded in 1924 by Howard W. Odum
- Oldest university-based interdisciplinary social science research institute in the U.S.
- *To help grow and lead a world-class social science research infrastructure at the University of North Carolina at Chapel Hill to ensure that researchers can conduct scientifically rigorous research that contributes to better lives of the citizens of North Carolina and the World…*

Research Data Management: Context is Critical

What are the components of your research data environment?

What would you need to reproduce your own research?

GETTING TO KNOW (AND LOVE) RESEARCH DATA

What are Data?

Research Data Management

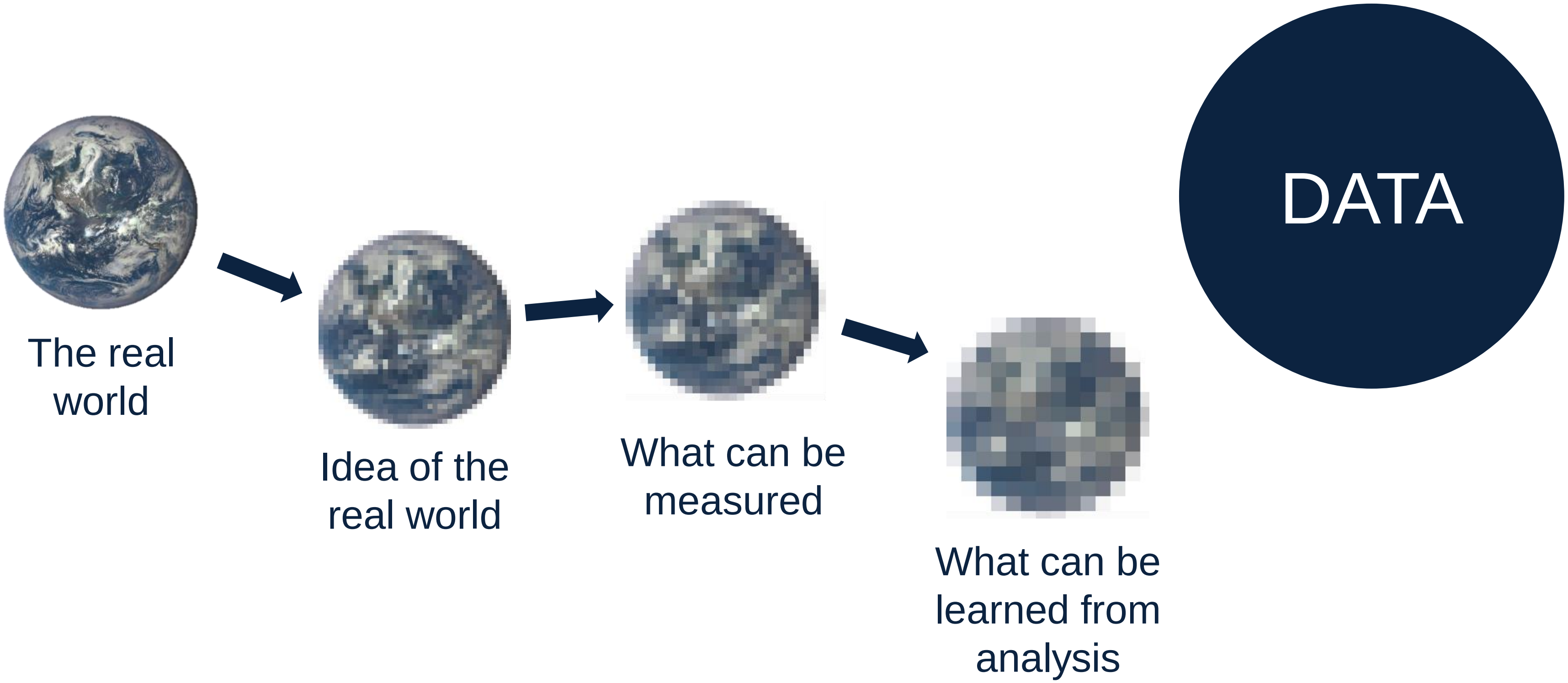
Data Documentation

Data Security

Archiving Data with Dataverse

www.digitalbevaring.dk

Odum Institute Services and Ongoing Projects



The real world

Idea of the real world

What can be measured

What can be learned from analysis

DATA

Little, J., & Zoss, A. (2014, September). *Basic data cleaning and analysis for data tables*. Webinar, Duke University Perkins Library. Retrieved from http://library.capture.duke.edu/Panopto/Pages/Viewer.aspx?id=9e7b8529-3566-4469-98f3-4e520f32b849

WHAT ARE DATA?

The world as DATA

- People
- Objects
- Places/Spaces
- Time
- Relationships
- Ideas/Concepts

WHAT ARE DATA?

- Telephone interview of 2,002 adults 18 or older
- Randomly selected youngest adult in household

Adapted photo by James Cridland available under CC BY 2.0 at https://flic.kr/p/Wd54U

Research Data Formats are Evolving

Increasing in size

Becoming more dynamic

More diverse

Harder to de-identify

More difficult to integrate with other data sources

Curation Processes

How do curation processes differ across file formats and disciplinary contexts?

WHY DATA MANAGEMENT?

Data management refers to activities that support long-term preservation, access, and use of data.

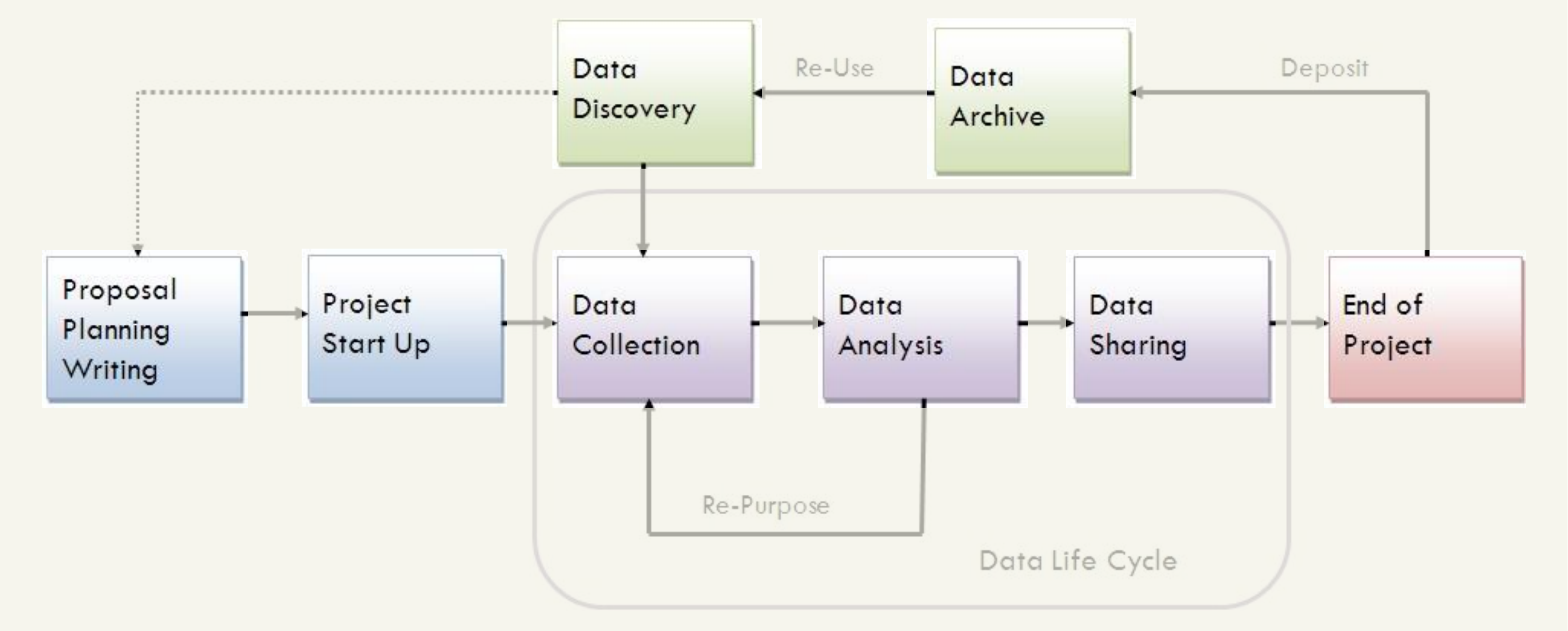
- Planning for data management
- Describing data
- Formatting data
- Storing and backing up data
- Anonymizing data
- Controlling access to data

WHY DATA MANAGEMENT?

- Data management makes it possible for other researchers to discover, interpret, and re-use data.
- Data management helps sustain the value of data by enabling others to verify and build upon published results.
- Data management facilitates long-term preservation of and access to data.

DATA MANAGEMENT PLANNING

Data management is most successful when data management practices are implemented throughout the research lifecycle.



Source: University of Virginia Library. (2014). Steps in the research lifecycle. Retrieved September 21, 2014, from http://dmconsult.library.virginia.edu/lifecycle/

DATA **DOCUMENTATION**

…**sufficient information** exists with which to understand, evaluate, and build upon a prior work if a third party could replicate the results **without any additional information from the author**.

King, G. (1995). Replication, replication. *PS: Political Science & Politics, 28*(3), 444–452. https://doi.org/10.2307/420301

Image source: http://harvardmagazine.com/2009/09/two-honored-with-university-professorships

DATA **DOCUMENTATION**

CODEBOOK

README

ANALYSIS CODE

DATA **DOCUMENTATION**

CODEBOOK

- Variable names + labels
- Value codes + labels
- Range of values
- Data type

DATA **DOCUMENTATION**

README

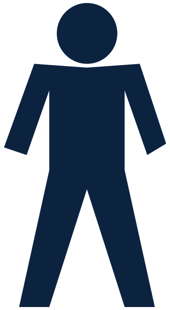
- Data collection methods
- Coding information
- Variable construction
- Dataset modifications
- Original data source

DATA **DOCUMENTATION**

ANALYSIS CODE

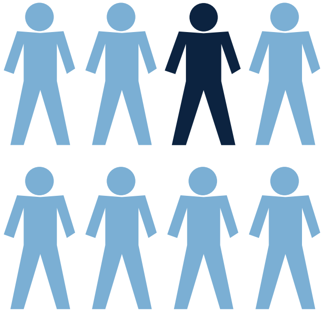
- Software version
- Commands
- Comment statements

DATA **SECURITY**



DIRECT IDENTIFIERS

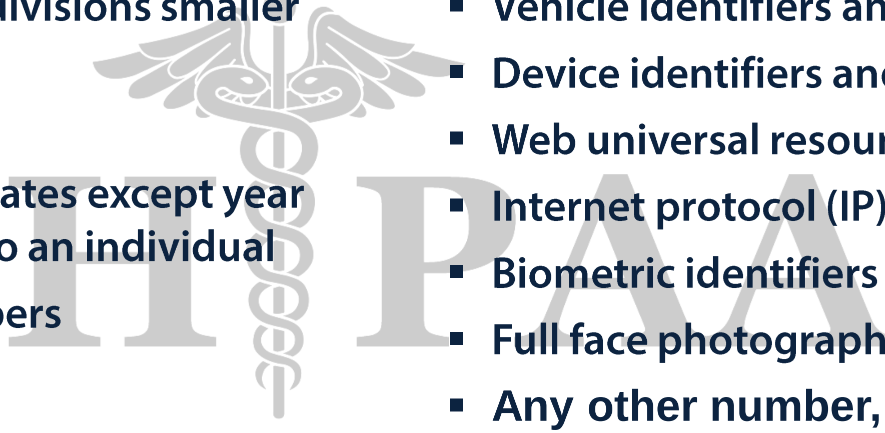
- Name
- Social Security Number
- Phone Number



INDIRECT IDENTIFIERS

- Race/Ethnicity
- Income
- Profession

DATA **SECURITY**



- Names
- Geographic subdivisions smaller than state
- Zip codes
- All elements of dates except year directly related to an individual
- Telephone numbers
- Fax numbers
- Email addresses
- Social Security numbers
- Medical record numbers
- Health plan beneficiary identifiers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers
- Device identifiers and serial numbers
- Web universal resource locators (URL)
- Internet protocol (IP) address numbers
- Biometric identifiers
- Full face photographic images
- Any other number, characteristic, or code that could be used by the researcher to identify the individual

DATA **SECURITY**

Date of birth
Gender
+ Zip code

87%
uniquely
identifiable

http://aboutmyinfo.org/

Sweeney, L. (2000). *Simple demographics often identify people uniquely* (Data Privacy Working Paper No. 3). Pittsburgh, PA: Carnegie Mellon University. Retrieved from http://dataprivacylab.org/projects/identifiability/paper1.pdf

Data security: threats and vulnerabilities

Sources of threat

- Natural
- Unintentional Human
- Intentional

Areas of vulnerability

- **Logical**: Data at rest in system, data in motion across networks, data being processed in applications
- **Physical**: Computer systems, network, and backups, disposal, media
- **Social**: social engineering, mistakes, insider threats

Altman, Micah. (2013). *Managing Confidential Data* [PowerPoint slides]. Retrieved from http://www.slideshare.net/drmaltman/altman-confidentialdata-v22mit?ref=http://informatics.mit.edu/classes/managing-confidential-data

Confidentiality & security across the research lifecycle

Identify potentially sensitive information

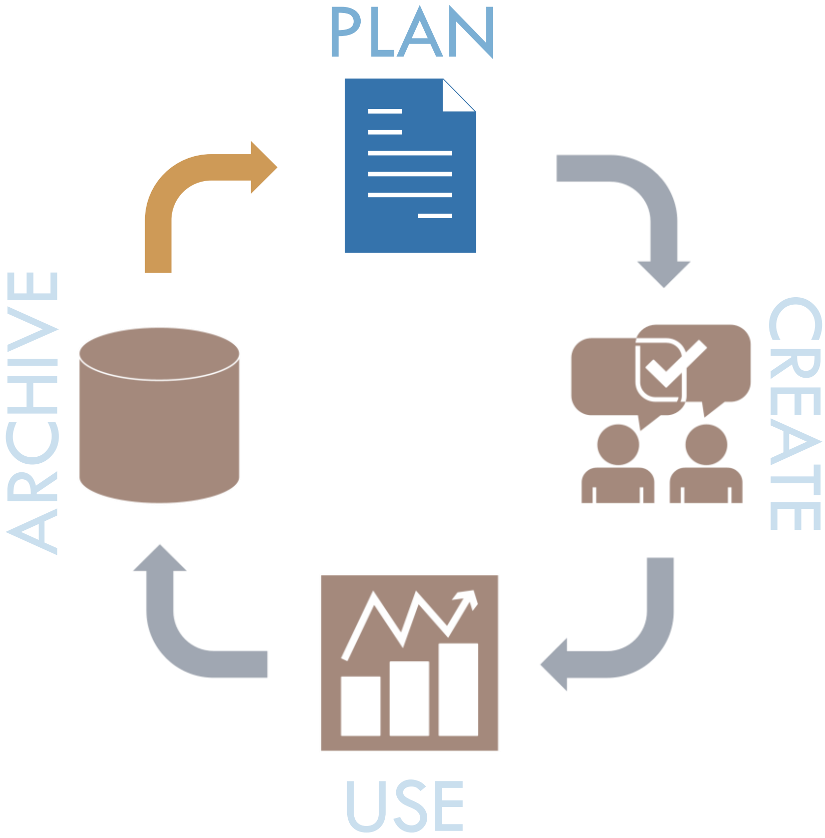
Review applicable laws

Design risk mitigation across data lifecycle

Reduce sensitivity of collected data

Plan for publication, dissemination, and reuse

Describe reuse plan in consent form

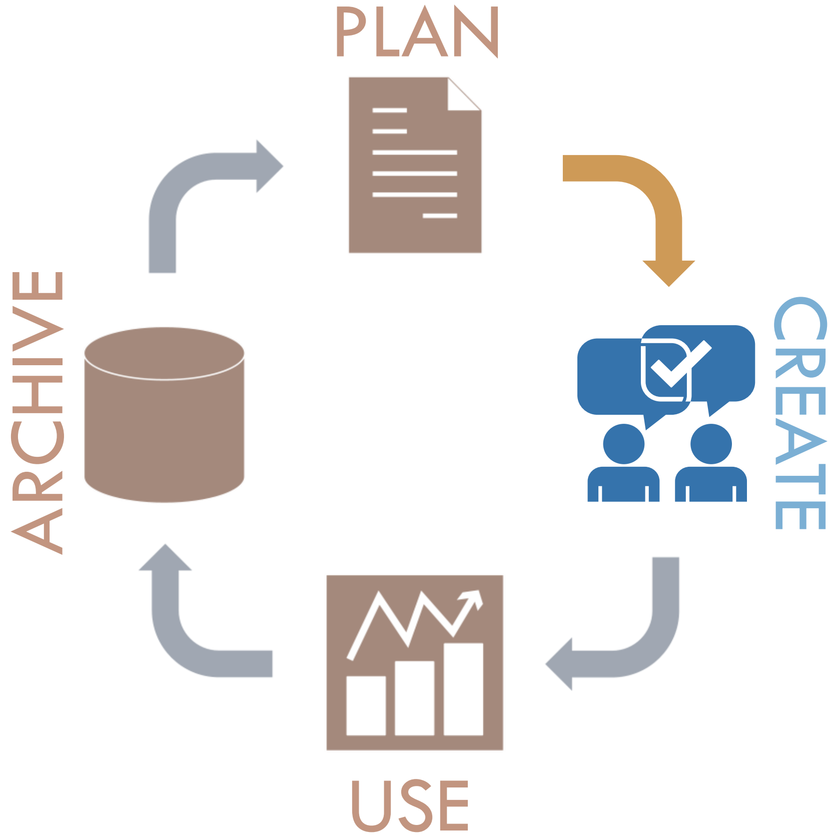


Confidentiality & security across the research lifecycle

Separate sensitive information in *collection*

Encrypt sensitive information in *transit*

Follow data security best practices



Confidentiality & security across the research lifecycle

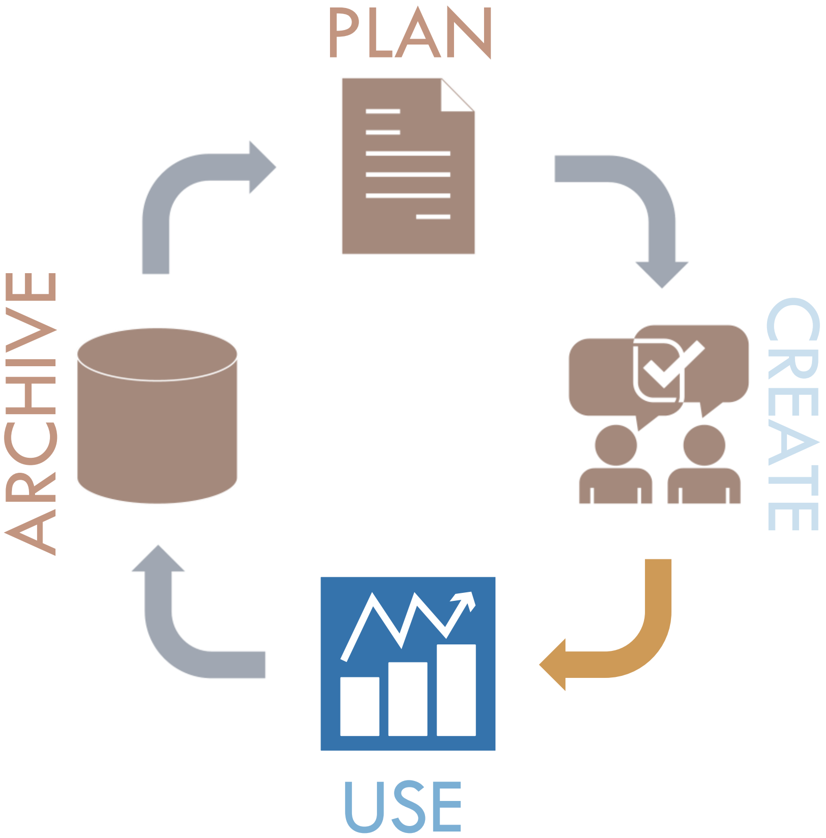
Protect sensitive information in *systems*

Desensitize information in *processing*

Monitor threats and vulnerabilities

Implement strategies for limiting disclosure risks

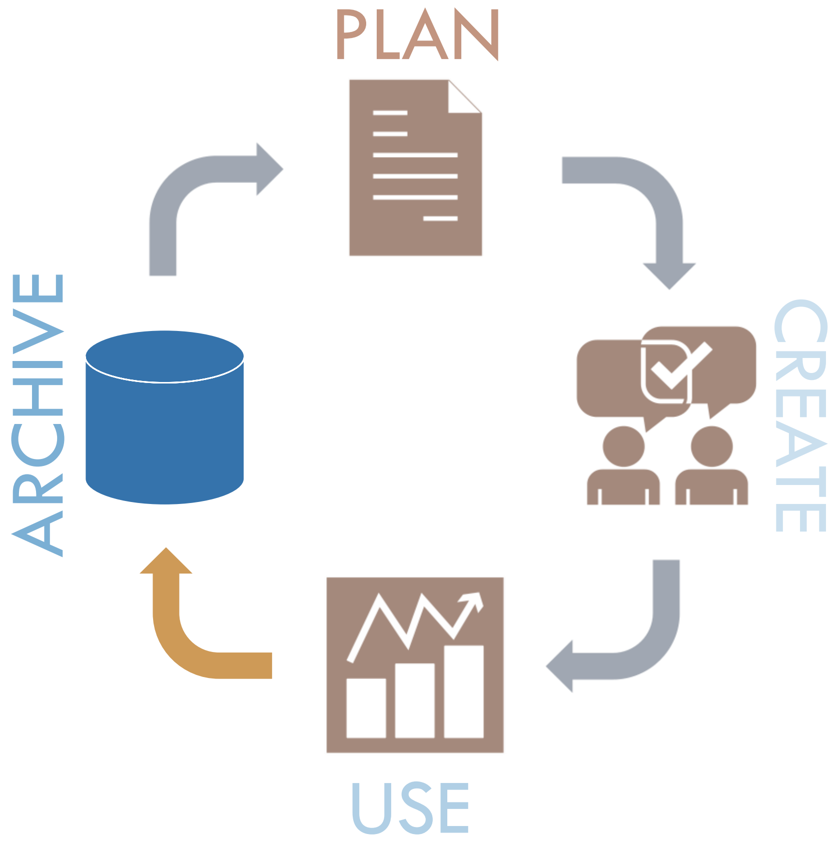
Review data for sensitive information prior to ingest into repository



Confidentiality & security across the research lifecycle

Deposit data in a trusted repository

Dispose of confidential data following best practices



Sharing safely: New approaches

Synthetic data

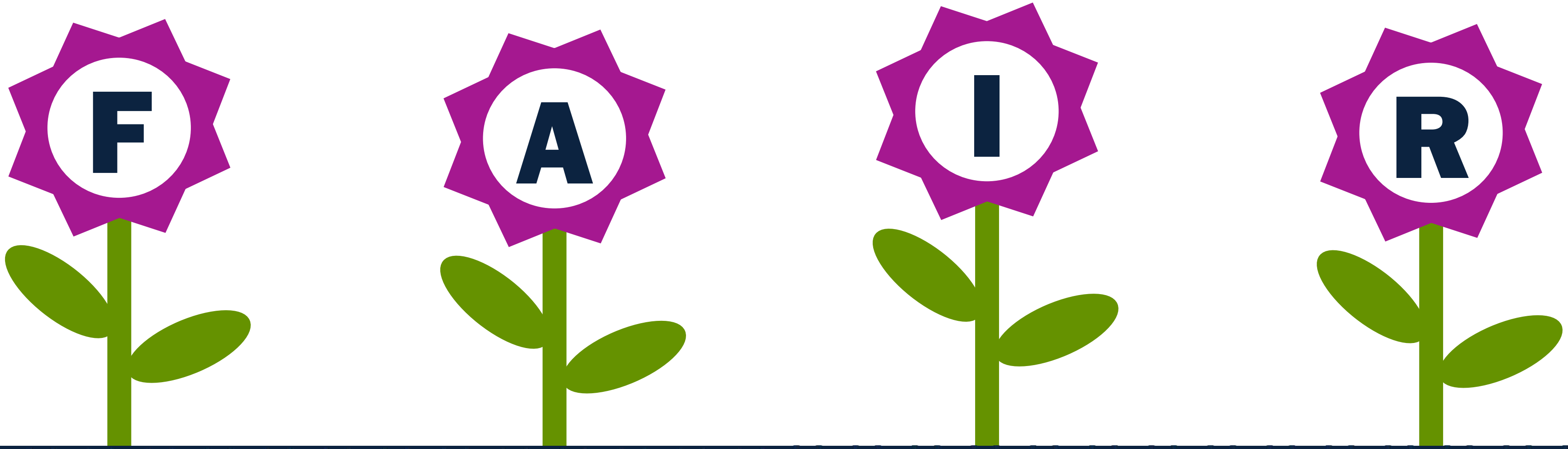
Differential Privacy

Database cryptography

Rules based data sharing tools

THE FAIR DATA PRINCIPLES

FORCE11. (2016). *Guiding principles for findable, accessible, interoperable and reusable data* (Publishing Version No. B1.0). Retrieved from https://www.force11.org/fairprinciples



Findable **A**ccessible **I**nteroperable **R**eusable

ARCHIVING DATA

A trusted digital repository is one whose mission is to provide reliable long-term access to managed digital resources to its designated community, now and in the future.

RLG/OCLC Working Group on Digital Archive Attributes. (2002). *Trusted digital repositories: Attributes and responsibilities* (An RLG-OCLC Report). Mountain View, CA: Research Libraries Group. Retrieved from http://www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf

TRUST Principles

FAIR defines the properties of data and metadata

TRUST describes the characteristics of data repositories that are responsible for managing and disseminating the data over a long period of time

FAIR data in repositories we **TRUST**

T - Transparency is achieved by providing publicly accessible evidence of the services that a repository can and can not offer.

R - Responsibility is a commitment to provide high (technical) quality data services.

U - User community is the focus on the uses and potential uses of the data and services offered.

S - Sustainability is the capability to support long-term data preservation and use.

T - Technology is the infrastructure and capabilities to support the repository operations.

FINDING A REPOSITORY

1. Is the repository reputable?
2. Will it take the data you want to deposit?
3. Will it be safe in legal terms?
4. Will the repository sustain the data value?
5. Will it support analysis and track data usage?

Whyte, A. (2015). Where to keep research data: DCC Checklist for evaluating data repositories (v.1.1). Edinburgh: Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides-checklists/where-keep-research-data

ARCHIVING DATA

Include documentation and metadata

Provide information to enable discovery and appropriate interpretation and reuse of the data

README FILE

CODEBOOK

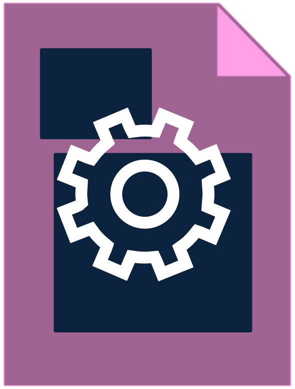
METADATA

ARCHIVING DATA

Include documentation and metadata

Provide information to enable discovery and appropriate interpretation and reuse of the data

ANALYSIS CODE

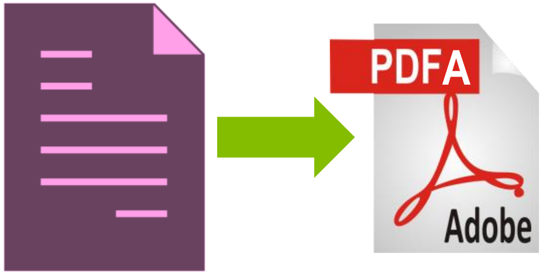


TIPS FOR ARCHIVING DATA

Archive data in open file formats

Use formats that support preservation, accessibility, and reuse of data

DOCUMENTATION



DATA FILES

.csv

.tab

.txt

TIPS FOR ARCHIVING DATA

Outline terms of use & apply a standard license

Enable informed reuse by clearly outlining how data can be accessed, used, and disseminated

DATA USE AGREEMENTS

EMBARGOES

CREATIVE COMMONS LICENSES

TIPS FOR ARCHIVING DATA

Resolve data ownership & sharing issues early

Discuss data sharing & archiving with collaborators, participants, and other stakeholders early in a project

Data citation

Provides stable access to data

Encourages acknowledgement and credit for data producers

Provides incentives for sharing data

Helios Herrera; Massimo Morelli; Salvatore Nunnari, 2015, "Replication data for: Turnout Across Democracies", http://dx.doi.org/10.7910/DVN/9TPNOT, Harvard Dataverse, V1 [UNF:6:NhH3rblEwGkIIbw9mygwCQ==]

Data citation

Joint Declaration of Data Citation Principles

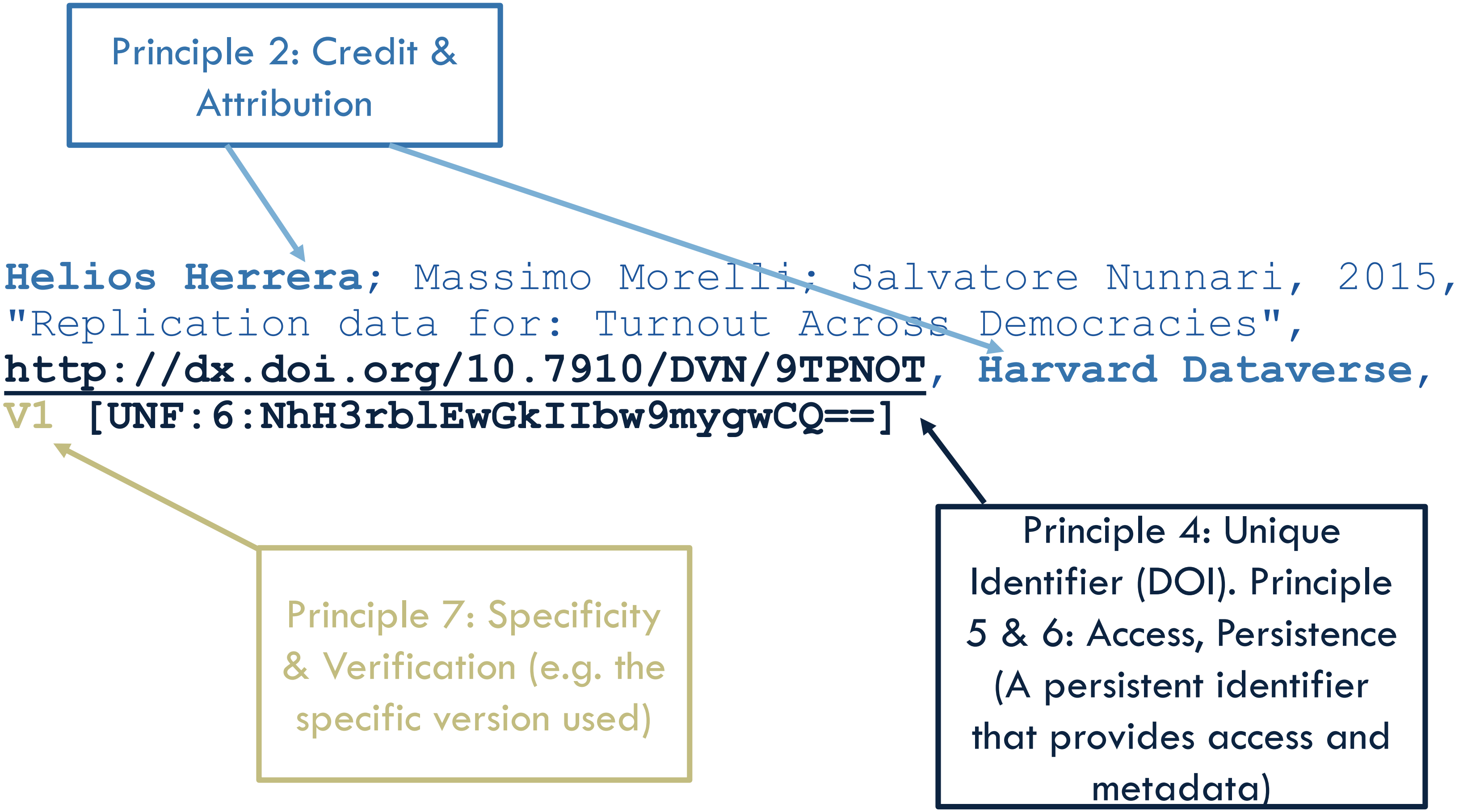
Synthesis of a number of groups and sponsored by Force 11

Facilitate the creation of citation practices that are both human understandable and machine-actionable

8 Data Citation Principles
1. Importance
2. Credit & Attribution
3. Evidence
4. Unique Identification
5. Access
6. Persistence
7. Specificity & Verifiability
8. Interoperability & Flexibility

Data Citation Synthesis Group: **Joint Declaration of Data Citation Principles**. Martone M. (ed.) San Diego CA: FORCE11; 2014 [https://www.force11.org/datacitation].

Data citation



Principle 2: Credit & Attribution

Helios Herrera; Massimo Morelli; Salvatore Nunnari, 2015, "Replication data for: Turnout Across Democracies", **http://dx.doi.org/10.7910/DVN/9TPNOT**, **Harvard Dataverse**, **V1** **[UNF:6:NhH3rblEwGkIIbw9mygwCQ==]**

Principle 7: Specificity & Verification (e.g. the specific version used)

Principle 4: Unique Identifier (DOI). Principle 5 & 6: Access, Persistence (A persistent identifier that provides access and metadata)

Data citation

Assign persistent identifiers (DOIs) to data

Supports simple & effective methods of data citation, discovery, and access

Ensures data can be located online

DataCite

FIND, ACCESS, AND REUSE DATA

The Data Sharing Problem

Different needs for archives, data libraries, researchers, journals, funding agencies…

We should preserve the data

We need persistent links

No publications without data

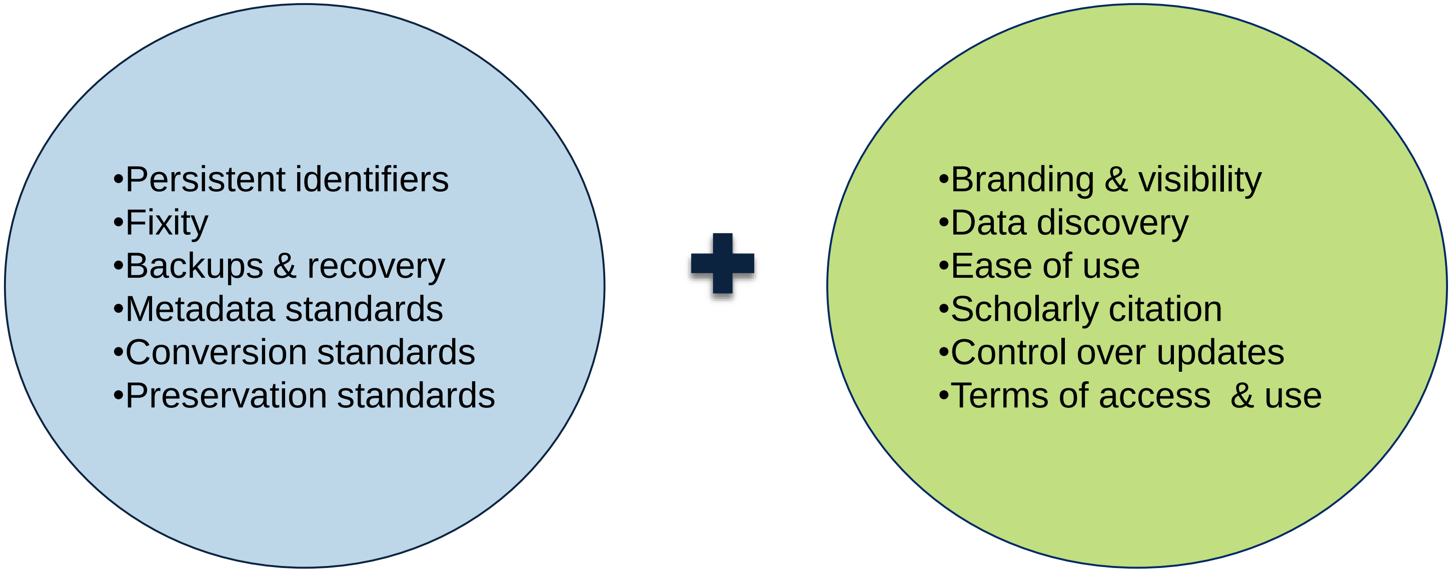
I want credit for my data

I need a Data Management Plan

Cross, M. *Why the Dataverse Network?* Available at: thedata.org

Odum’s Solution

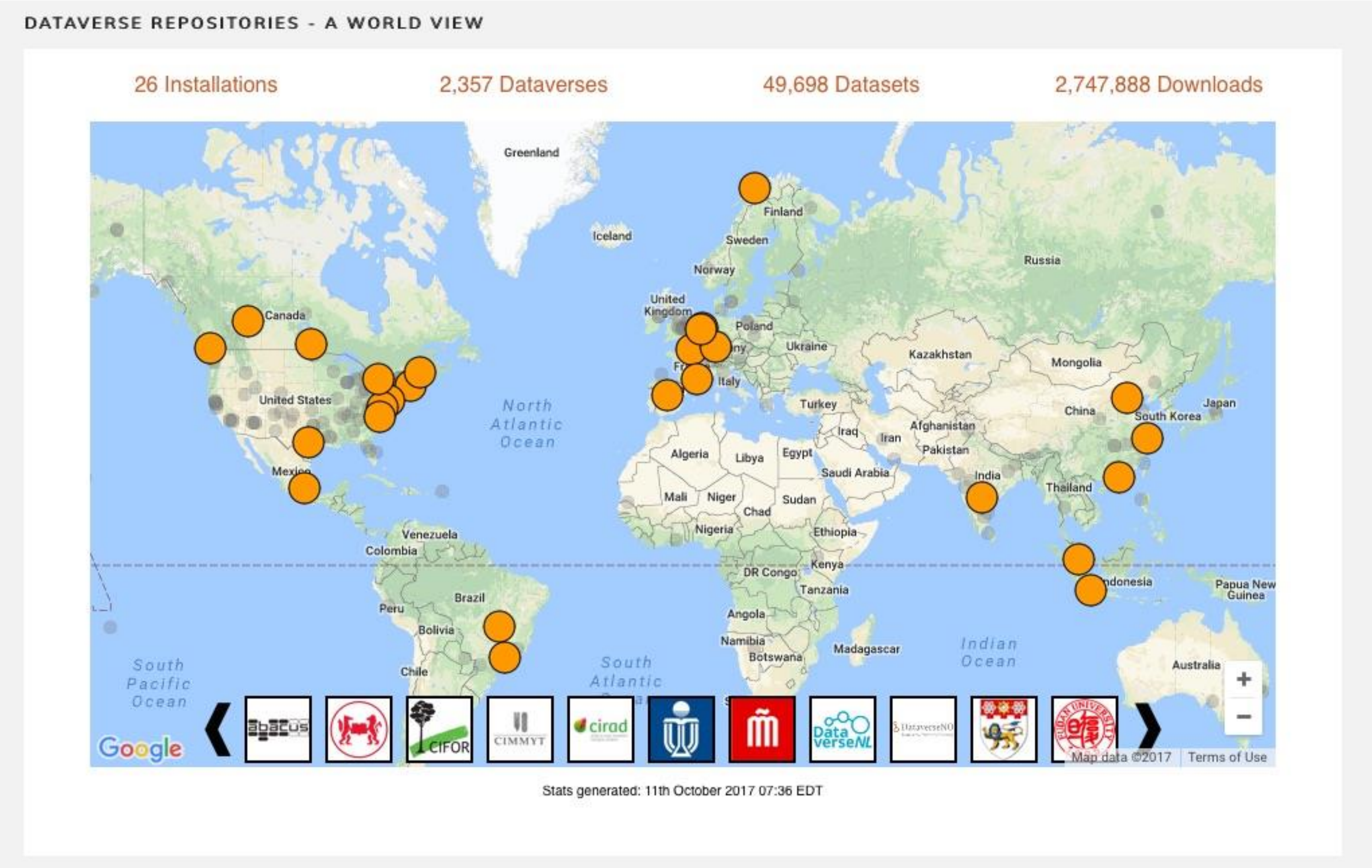
Dataverse: **centralized** professional archiving with **distributed** control and recognition



- Persistent identifiers
- Fixity
- Backups & recovery
- Metadata standards
- Conversion standards
- Preservation standards

- Branding & visibility
- Data discovery
- Ease of use
- Scholarly citation
- Control over updates
- Terms of access & use

Cross, M. *Why the Dataverse Network?* Available at: thedata.org



DATAVERSE REPOSITORIES - A WORLD VIEW

26 Installations

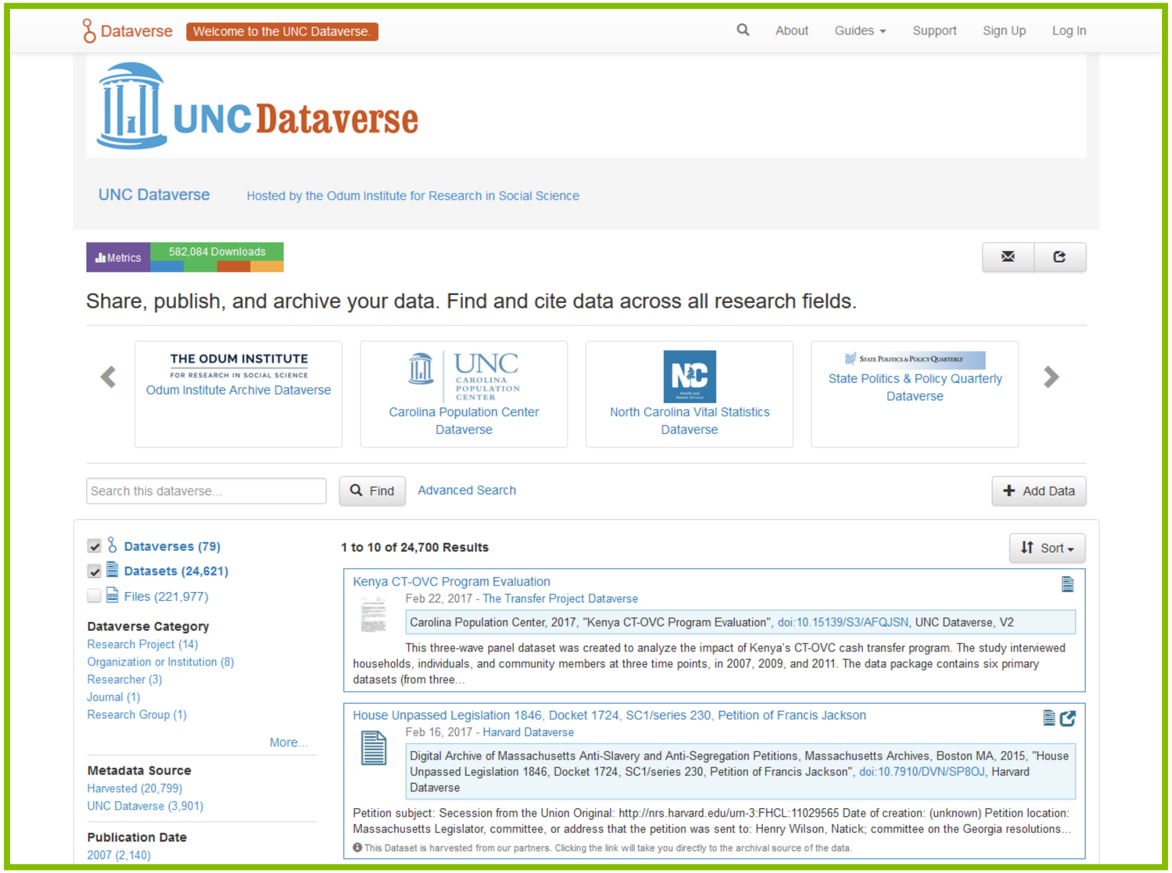
2,357 Dataverses

49,698 Datasets

2,747,888 Downloads

Stats generated: 11th October 2017 07:36 EDT

THE DATAVERSE PROJECT



Dataverse Welcome to the UNC Dataverse.

About Guides Support Sign Up Log In

UNC Dataverse

UNC Dataverse Hosted by the Odum Institute for Research in Social Science

Metrics 582,084 Downloads

Share, publish, and archive your data. Find and cite data across all research fields.

THE ODUM INSTITUTE FOR RESEARCH IN SOCIAL SCIENCE
Odum Institute Archive Dataverse

UNC CAROLINA POPULATION CENTER
Carolina Population Center Dataverse

NC
North Carolina Vital Statistics Dataverse

State Politics & Policy Quarterly
State Politics & Policy Quarterly Dataverse

Search this dataverse... Find Advanced Search Add Data

☑ Dataverses (79)
☑ Datasets (24,621)
□ Files (221,977)

Dataverse Category
Research Project (14)
Organization or Institution (8)
Researcher (3)
Journal (1)
Research Group (1)
More...

Metadata Source
Harvested (20,799)
UNC Dataverse (3,901)

Publication Date
2007 (2,140)

1 to 10 of 24,700 Results Sort

Kenya CT-OVC Program Evaluation
Feb 22, 2017 - The Transfer Project Dataverse
Carolina Population Center, 2017, "Kenya CT-OVC Program Evaluation", doi:10.15139/S3/AFQJSN, UNC Dataverse, V2
This three-wave panel dataset was created to analyze the impact of Kenya's CT-OVC cash transfer program. The study interviewed households, individuals, and community members at three time points, in 2007, 2009, and 2011. The data package contains six primary datasets (from three...

House Unpassed Legislation 1846, Docket 1724, SC1/series 230, Petition of Francis Jackson
Feb 16, 2017 - Harvard Dataverse
Digital Archive of Massachusetts Anti-Slavery and Anti-Segregation Petitions, Massachusetts Archives, Boston MA, 2015, "House Unpassed Legislation 1846, Docket 1724, SC1/series 230, Petition of Francis Jackson", doi:10.7910/DVN/SP8OJ, Harvard Dataverse
Petition subject: Secession from the Union Original: http://nrs.harvard.edu/urn-3:FHCL:11029565 Date of creation: (unknown) Petition location: Massachusetts Legislator, committee, or address that the petition was sent to: Henry Wilson, Natick; committee on the Georgia resolutions...
This Dataset is harvested from our partners. Clicking the link will take you directly to the archival source of the data.

https://dataverse.unc.edu/

THE DATAVERSE PROJECT

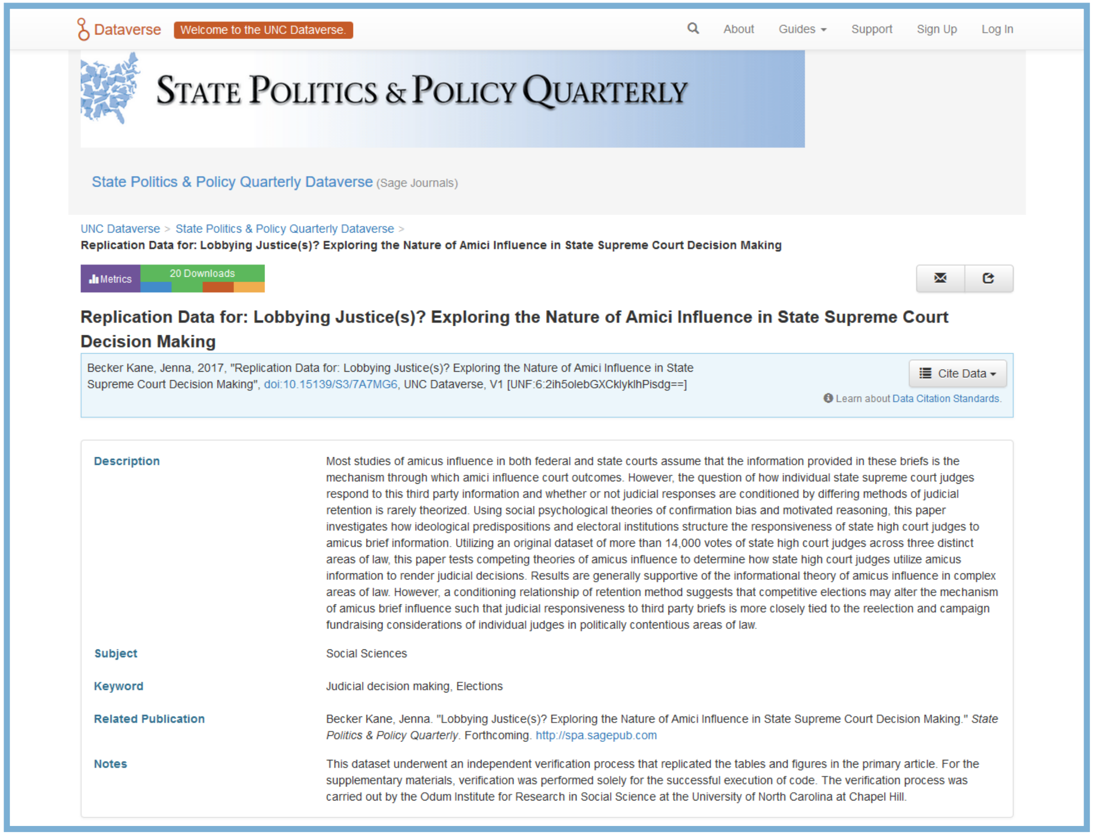
- Open source web application for publishing, citing, analyzing, and preserving research data
- Data sharing and archiving with control and recognition for data producers
- Rich data support for certain file formats
- Supports data management standards and best practices

The Dataverse Project

THE DATAVERSE PROJECT

Archival Record

- Standardized DDI metadata
- Formal citation
- Persistent identification



State Politics & Policy Quarterly Dataverse (Sage Journals)

UNC Dataverse > State Politics & Policy Quarterly Dataverse >
Replication Data for: Lobbying Justice(s)? Exploring the Nature of Amici Influence in State Supreme Court Decision Making

Replication Data for: Lobbying Justice(s)? Exploring the Nature of Amici Influence in State Supreme Court Decision Making

Becker Kane, Jenna, 2017, "Replication Data for: Lobbying Justice(s)? Exploring the Nature of Amici Influence in State Supreme Court Decision Making", doi:10.15139/S3/7A7MG6, UNC Dataverse, V1 [UNF:6:2ih5olebGXCklyklhPisdg==]

Learn about Data Citation Standards.

Description

Most studies of amicus influence in both federal and state courts assume that the information provided in these briefs is the mechanism through which amici influence court outcomes. However, the question of how individual state supreme court judges respond to this third party information and whether or not judicial responses are conditioned by differing methods of judicial retention is rarely theorized. Using social psychological theories of confirmation bias and motivated reasoning, this paper investigates how ideological predispositions and electoral institutions structure the responsiveness of state high court judges to amicus brief information. Utilizing an original dataset of more than 14,000 votes of state high court judges across three distinct areas of law, this paper tests competing theories of amicus influence to determine how state high court judges utilize amicus information to render judicial decisions. Results are generally supportive of the informational theory of amicus influence in complex areas of law. However, a conditioning relationship of retention method suggests that competitive elections may alter the mechanism of amicus brief influence such that judicial responsiveness to third party briefs is more closely tied to the reelection and campaign fundraising considerations of individual judges in politically contentious areas of law.

Subject

Social Sciences

Keyword

Judicial decision making, Elections

Related Publication

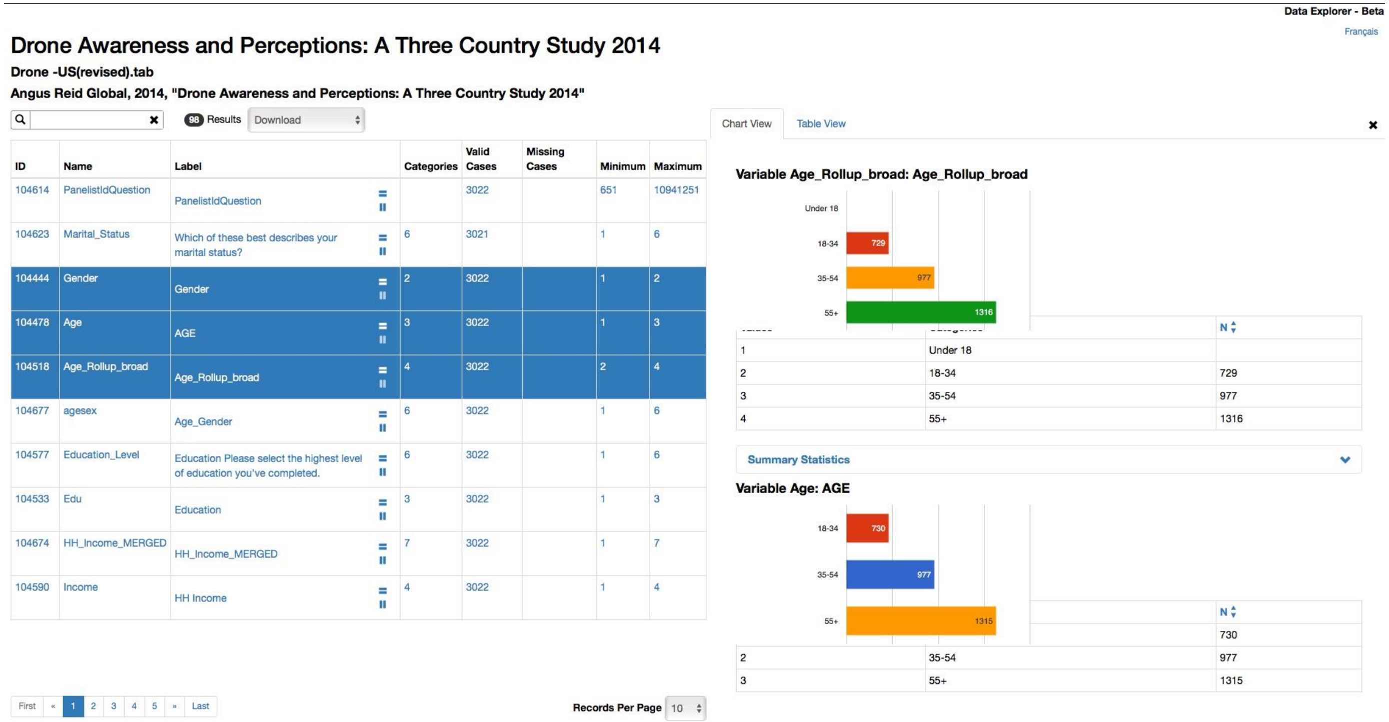
Becker Kane, Jenna. "Lobbying Justice(s)? Exploring the Nature of Amici Influence in State Supreme Court Decision Making." *State Politics & Policy Quarterly*. Forthcoming. http://spa.sagepub.com

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.

https://dataverse.unc.edu/

External Analysis Tools via Dataverse API



Data Explorer - Beta

Français

Drone Awareness and Perceptions: A Three Country Study 2014

Drone -US(revised).tab

Angus Reid Global, 2014, "Drone Awareness and Perceptions: A Three Country Study 2014"

98 Results Download

ID	Name	Label	Categories	Valid Cases	Missing Cases	Minimum	Maximum
104614	PanelistIdQuestion	PanelistIdQuestion		3022		651	10941251
104623	Marital_Status	Which of these best describes your marital status?	6	3021		1	6
104444	Gender	Gender	2	3022		1	2
104478	Age	AGE	3	3022		1	3
104518	Age_Rollup_broad	Age_Rollup_broad	4	3022		2	4
104677	agesex	Age_Gender	6	3022		1	6
104577	Education_Level	Education Please select the highest level of education you've completed.	6	3022		1	6
104533	Edu	Education	3	3022		1	3
104674	HH_Income_MERGED	HH_Income_MERGED	7	3022		1	7
104590	Income	HH Income	4	3022		1	4

Records Per Page 10

Variable Age_Rollup_broad: Age_Rollup_broad

		N
1	Under 18	
2	18-34	729
3	35-54	977
4	55+	1316

Summary Statistics

Variable Age: AGE

		N
		730
2	35-54	977
3	55+	1315

Moving beyond social science

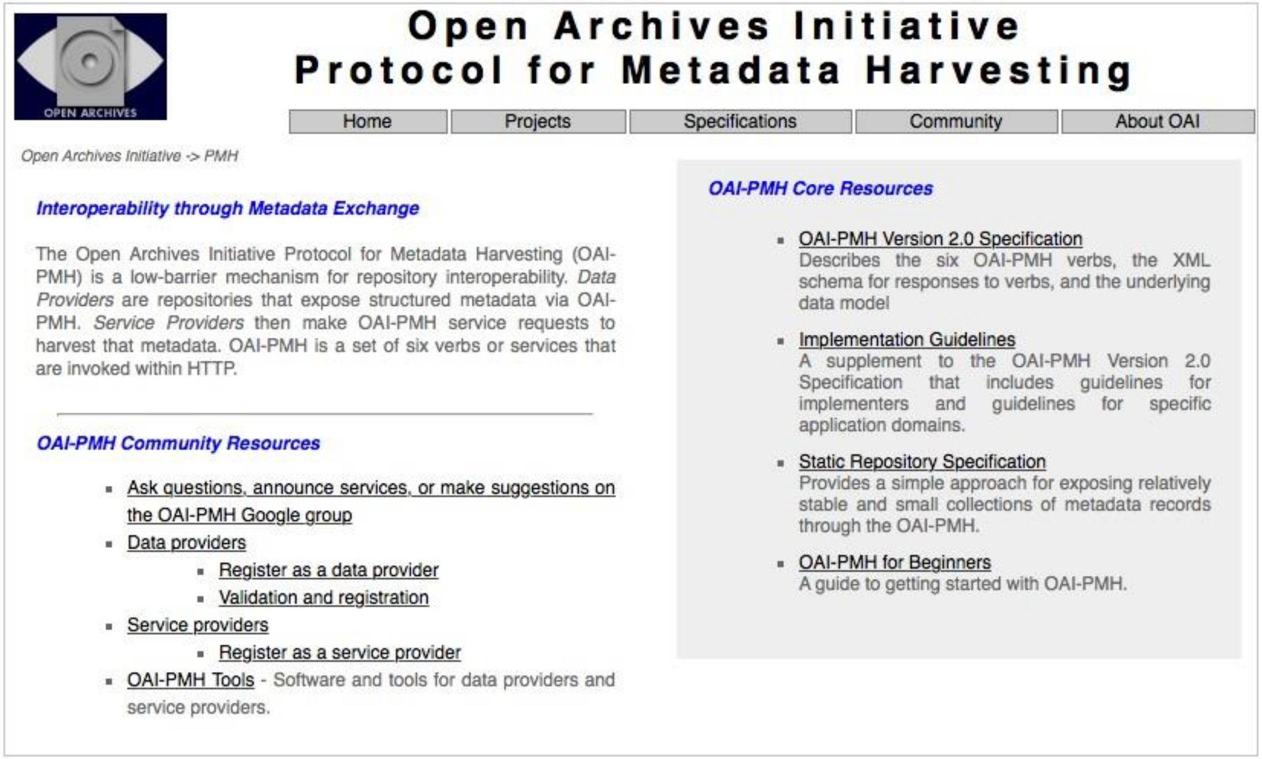
Dataverse Network is cross-disciplinary.

We are expanding the study metadata and building communities of interested groups:

- dataverse-community@googlegroups.com

Cross, M. *Why the Dataverse Network?* Available at: thedata.org

Support for metadata sharing



Open Archives Initiative
Protocol for Metadata Harvesting

Open Archives Initiative -> PMH

Interoperability through Metadata Exchange

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a low-barrier mechanism for repository interoperability. *Data Providers* are repositories that expose structured metadata via OAI-PMH. *Service Providers* then make OAI-PMH service requests to harvest that metadata. OAI-PMH is a set of six verbs or services that are invoked within HTTP.

OAI-PMH Community Resources

- Ask questions, announce services, or make suggestions on the OAI-PMH Google group
- Data providers
  - Register as a data provider
  - Validation and registration
- Service providers
  - Register as a service provider
- OAI-PMH Tools - Software and tools for data providers and service providers.

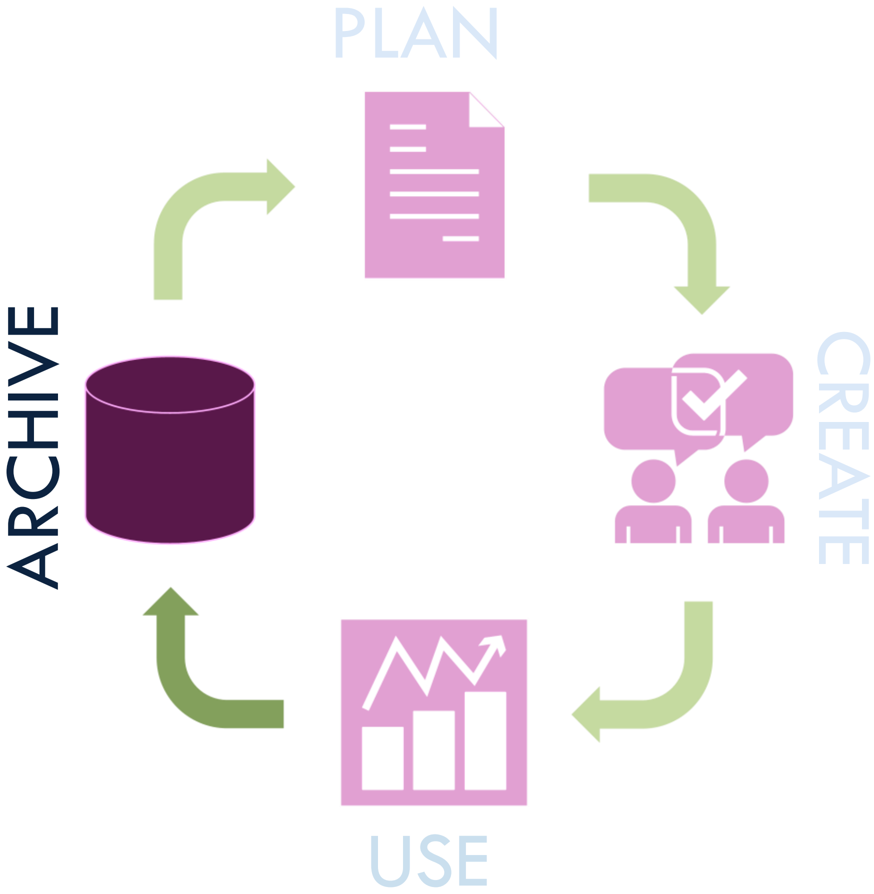
OAI-PMH Core Resources

- OAI-PMH Version 2.0 Specification
  Describes the six OAI-PMH verbs, the XML schema for responses to verbs, and the underlying data model
- Implementation Guidelines
  A supplement to the OAI-PMH Version 2.0 Specification that includes guidelines for implementers and guidelines for specific application domains.
- Static Repository Specification
  Provides a simple approach for exposing relatively stable and small collections of metadata records through the OAI-PMH.
- OAI-PMH for Beginners
  A guide to getting started with OAI-PMH.

THE H. W. ODUM INSTITUTE FOR RESEARCH IN SOCIAL SCIENCE

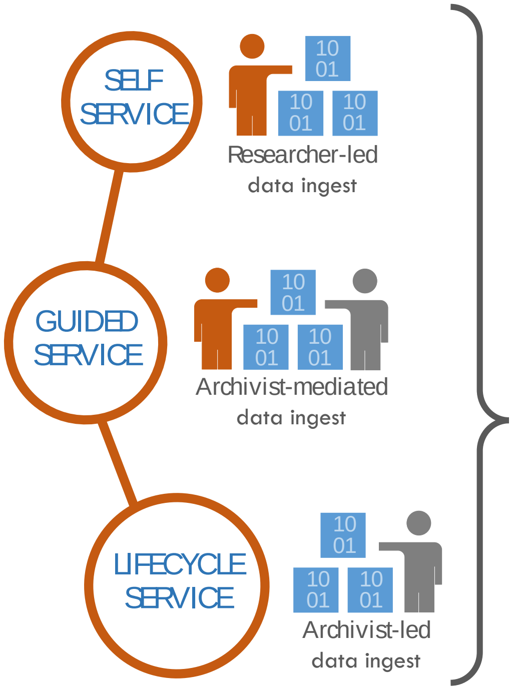
SERVICES ACROSS THE RESEARCH LIFEC

DATA ARCHIVE

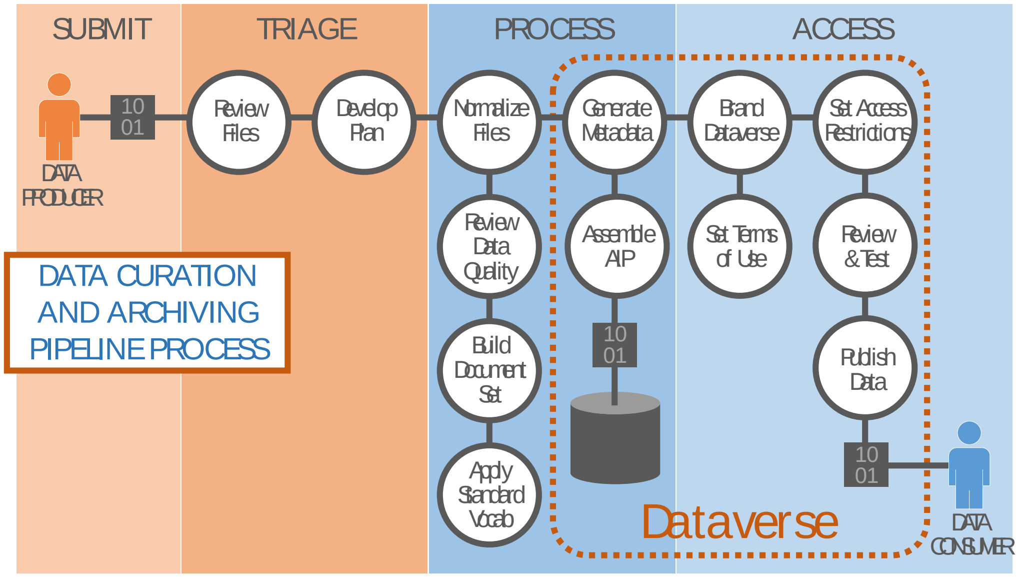


THE ODUM INSTITUTE FOR RESEARCH IN SOCIAL SCIENCE | THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

DATA CURATION AND ARCHIVING SERVICES POWERED BY Dataverse

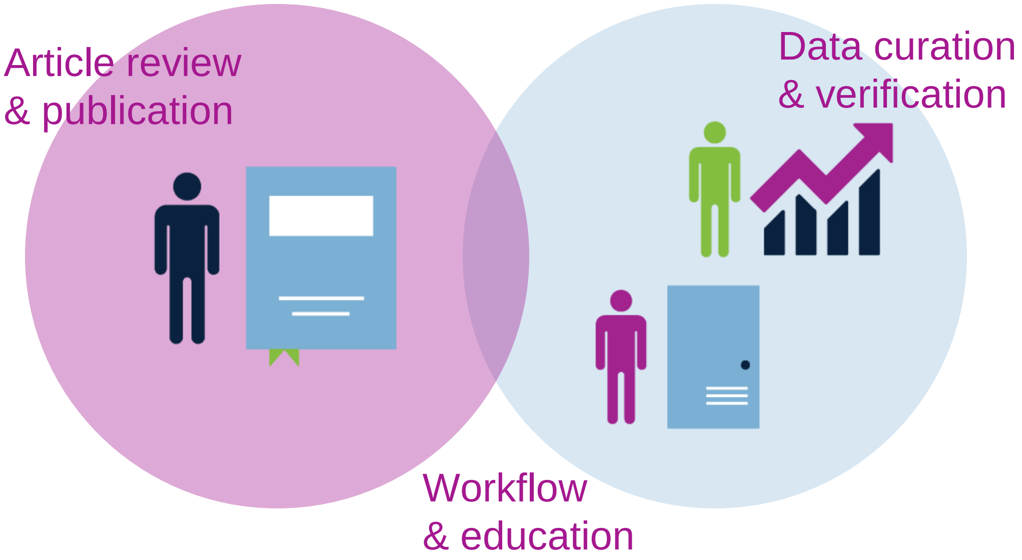


The Odum Institute for Research in Social Science offers three tiers of data curation and archiving services, which are based on the degree to which Odum Institute archivists provide guidance for or execute tasks in the data curation and archiving pipeline process.



http://www.odum.unc.edu | odumarchive@unc.edu

Data Verification Service

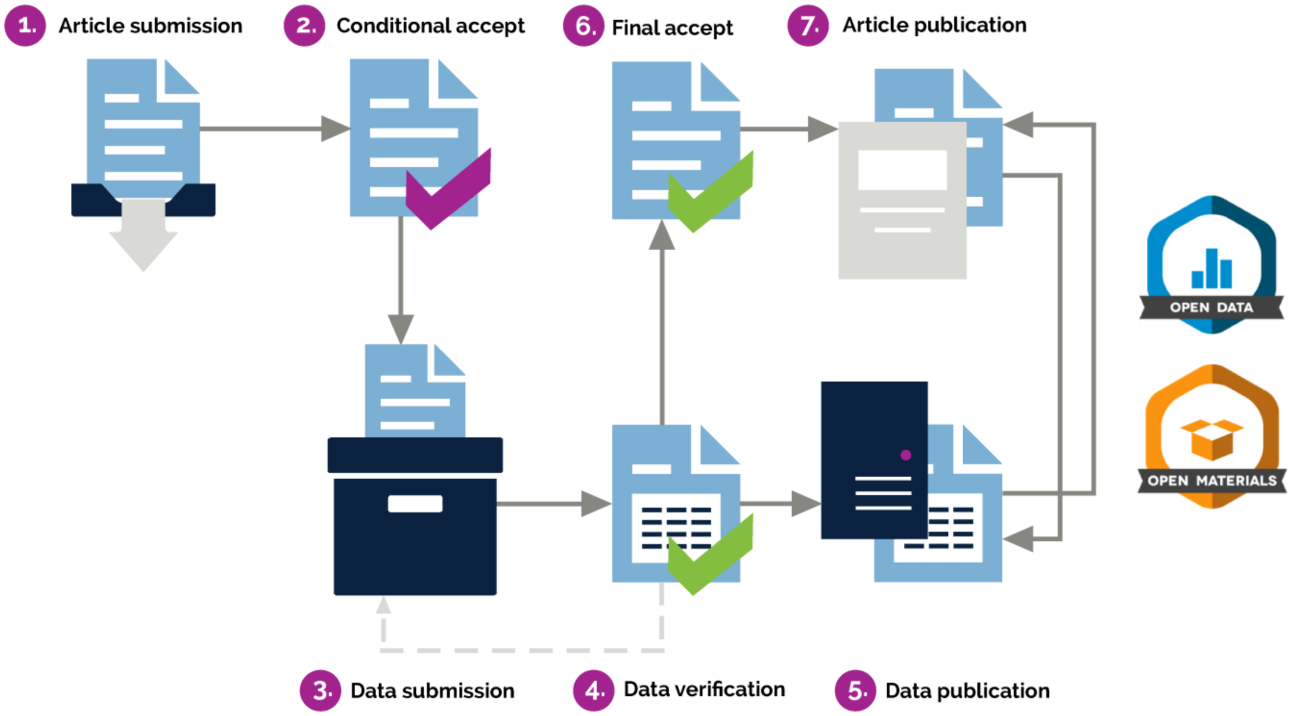


Article review
& publication

Data curation
& verification

Workflow
& education

Data verification



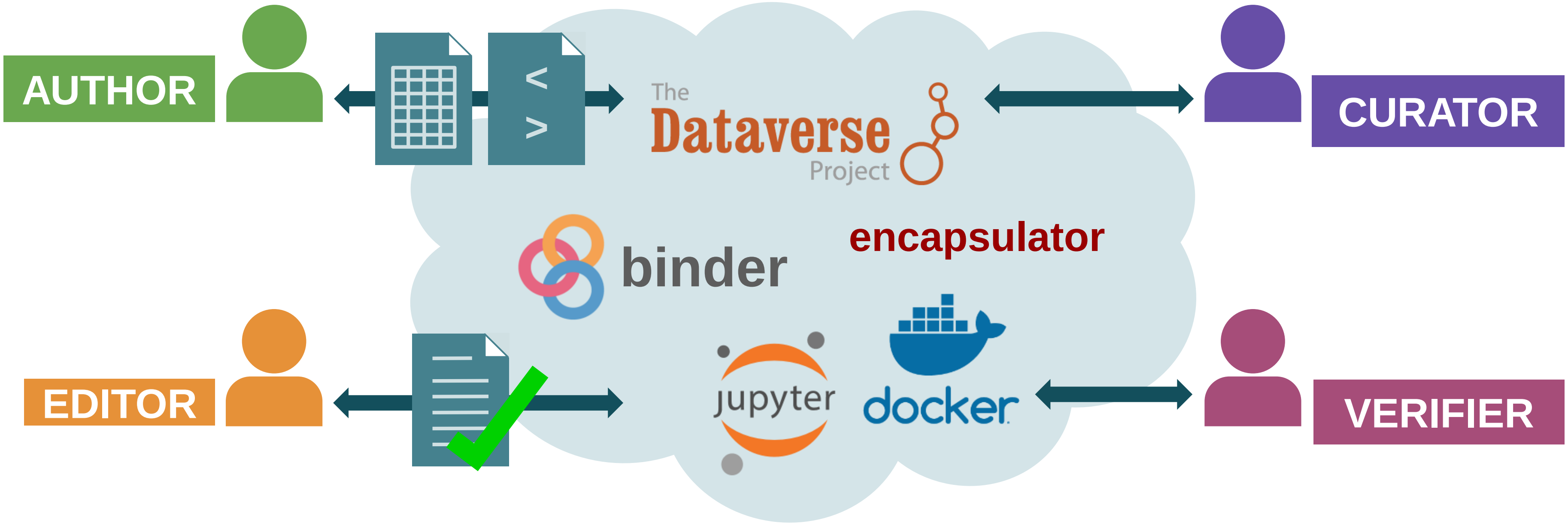
The Confirmable Reproducible Research (CoRe2) Environment

Linking Tools to Promote Computational Reproducibility

Alfred P. Sloan
FOUNDATION

Support for this research was provided by the Alfred P. Sloan Foundation (2018-11121). The views expressed here do not necessarily reflect the views of the Foundation.

Confirmable Reproducible Research (CoRe2) Environment

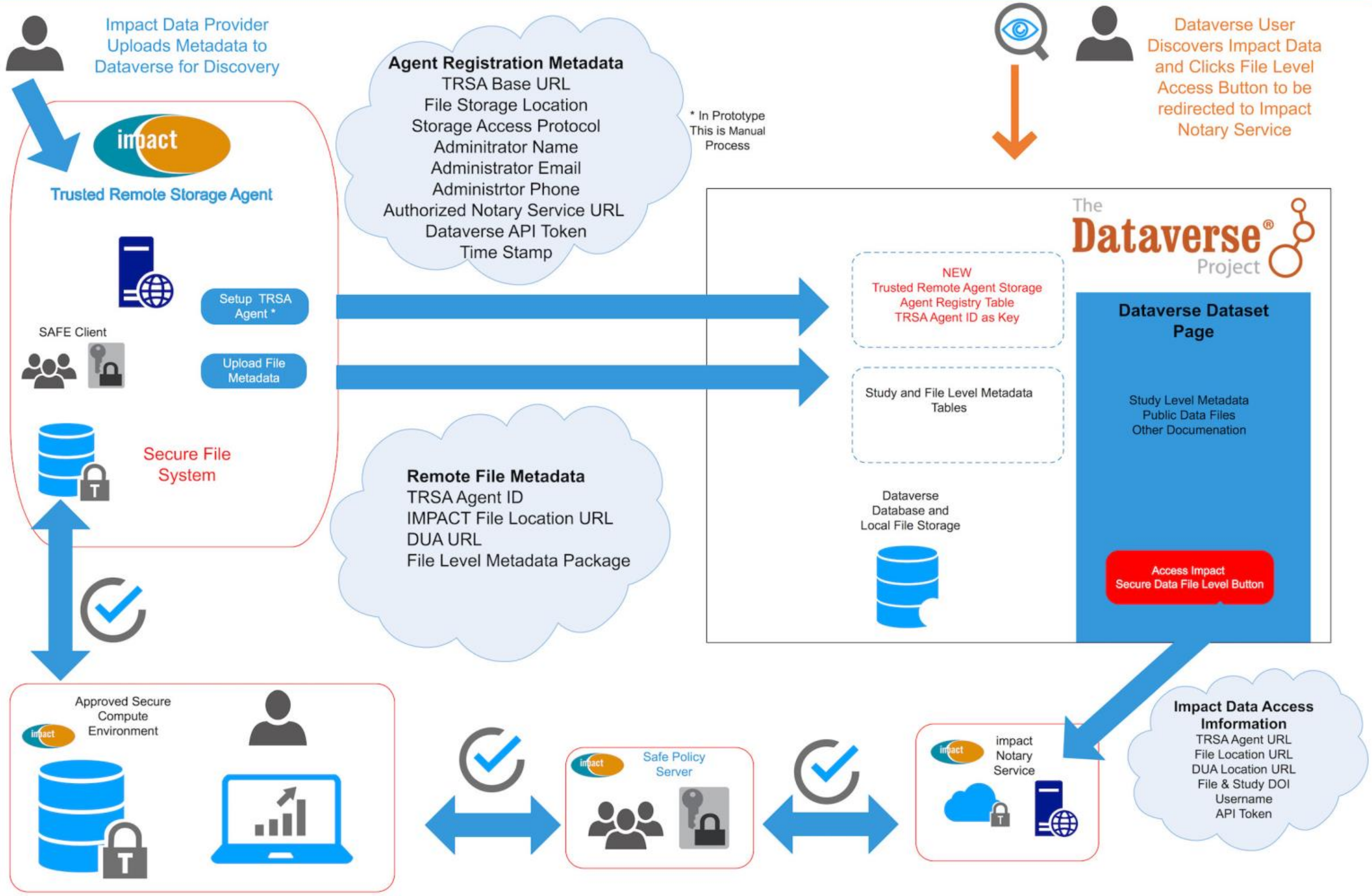


Impact Project Overview

- **Challenges:**
  - Social science and many other data-oriented disciplines depend on data belonging to multiple stakeholders
  - Governed by a variety of use policies
  - Multi-institutional research requires cooperative analysis
  - Need to satisfy the privacy concerns of the owners while producing interesting research outcomes by analyzing data

- **Goal:** to enable cooperative processing across the stakeholder-owned datasets, while respecting the privacy policies of the individual owners, and to provide a model for collaboration that could be readily used by other institutions.

IMPACT TRSA User Workflow



Impact Data Provider Uploads Metadata to Dataverse for Discovery

impact

Trusted Remote Storage Agent

Setup TRSA Agent *

Upload File Metadata

SAFE Client

Secure File System

Agent Registration Metadata
TRSA Base URL
File Storage Location
Storage Access Protocol
Admintrator Name
Administrator Email
Administrtor Phone
Authorized Notary Service URL
Dataverse API Token
Time Stamp

* In Prototype This is Manual Process

Remote File Metadata
TRSA Agent ID
IMPACT File Location URL
DUA URL
File Level Metadata Package

Dataverse User Discovers Impact Data and Clicks File Level Access Button to be redirected to Impact Notary Service

The Dataverse® Project

NEW
Trusted Remote Agent Storage Agent Registry Table
TRSA Agent ID as Key

Study and File Level Metadata Tables

Dataverse Database and Local File Storage

Dataverse Dataset Page

Study Level Metadata
Public Data Files
Other Documenation

Access Impact Secure Data File Level Button

Impact Data Access Information
TRSA Agent URL
File Location URL
DUA Location URL
File & Study DOI
Username
API Token

impact Notary Service

Safe Policy Server

Approved Secure Compute Environment

Thank You

CONNECT WITH THE ODUM INSTITUT

Jonathan Crabtree
Jonathan_Crabtree@unc.edu

The Odum Institute
http://www.odum.unc.edu

@Odum_Institute

The Odum Institute