MANAGING AND ARCHIVING DATA THROUGHOUT THE RESEARCH LIFECYCLE

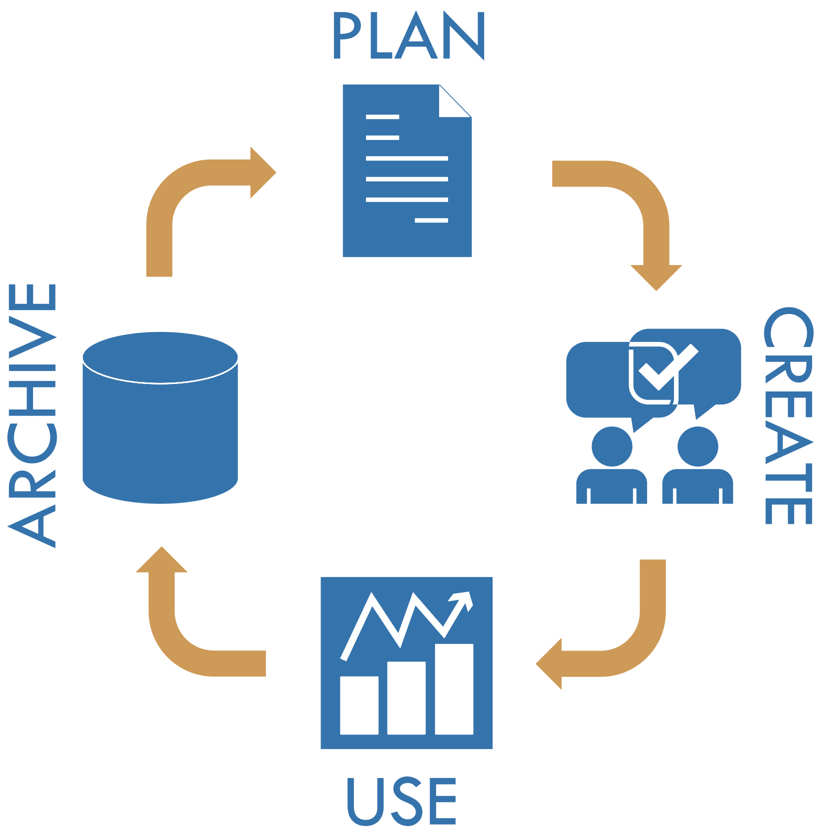
THOMAS M. CARSEY

Director carsey@unc.edu

UNC
THE ODUM INSTITUTE

THE ODUM INSTITUTE FOR RESEARCH IN SOCIAL SCIENCE
228 DAVIS LIBRARY, CB# 3355
UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
WWW.ODUM.UNC.EDU

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



DATA ARCHIVE

GRANT SERVICES

FACILITIES & LABS

DATA COLLECTION & ANALYSIS

CONSULTATIONS

TRAINING & EDUCATION

"DATA CURATION"

Data curation is "the active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education, which includes appraisal and selection, representation, and organization of these data for access and use over time."

Shreeves, S. L., & Cragin, M. H. (2008). Introduction: Institutional repositories: Current state and future. *Library Trends*, *57*(2), 89–97. http://doi.org/10.1353/lib.0.0037

WHY DATA MANAGEMENT?

- Data management improves the quality and integrity of your own research.
- Data management makes it possible for other researchers to discover, interpret, and reuse 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.

WHY DATA MANAGEMENT?

- A growing number of funding agencies, journal publishers, and institutions require it.

PLOS

NSF

NATIONAL INSTITUTES OF HEALTH

NATIONAL ENDOWMENT FOR THE HUMANITIES

WHY DATA MANAGEMENT?

- A growing number of funding agencies, journal publishers, and institutions require it.

VOLUME 60 NUMBER 1 JANUARY 2016

AJPS

AMERICAN JOURNAL of POLITICAL SCIENCE

Committed to significant advances in knowledge and understanding of politics, governance, and citizenship, and to the public value of political science research

Official Publication of the Midwest Political Science Association
www.mpsanet.org • www.ajps.org

The corresponding author of a manuscript…must provide replication materials that are sufficient to enable interested researchers to reproduce all of the analytic results that are reported in the text and supporting materials…the replication materials will be verified to confirm that they do, in fact, reproduce the analytic results reported in the article.

Jacoby, W., & Lupton, R. (2015). AJPS Replication and verification policy. Retrieved from https://ajps.org/ajps-replication-policy/

KEY STAKEHOLDERS

- Researchers
- Institutions
- Data repositories
- Secondary Users
- Funders
- Editors and Publishers

Lyon, L. (2007). *Dealing with data: Roles, rights, responsibilities, and relationships* (Consultancy Report). UK: UKOLN, University of Bath. Retrieved from http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf.

KEY STAKEHOLDERS:
EDITOR AND PUBLISHER

ROLE

Maintain the integrity of the scientific record

RIGHTS

RESPONSIBILITIES

Lyon, L. (2007). *Dealing with data: Roles, rights, responsibilities, and relationships* (Consultancy Report). UK: UKOLN, University of Bath. Retrieved from http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf.

KEY STAKEHOLDERS:
EDITOR AND PUBLISHER

ROLE

RIGHTS

RESPONSIBILITIES

- To expect data are available to support published results
- To request pre-publication data deposit in a data repository

Lyon, L. (2007). *Dealing with data: Roles, rights, responsibilities, and relationships* (Consultancy Report). UK: UKOLN, University of Bath. Retrieved from http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf.

KEY STAKEHOLDERS:
EDITOR AND PUBLISHER

ROLE

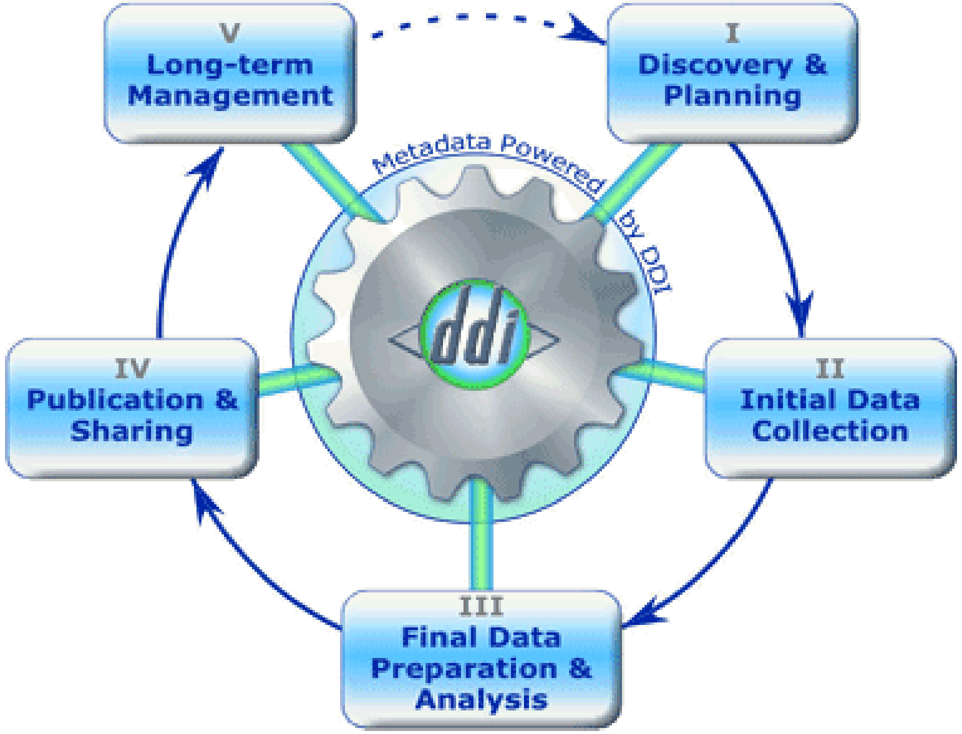
RIGHTS

RESPONSIBILITIES

- Engage stakeholders in the development of publication standards
- Link to data to support publication standards
- Monitor and enforce data management standards

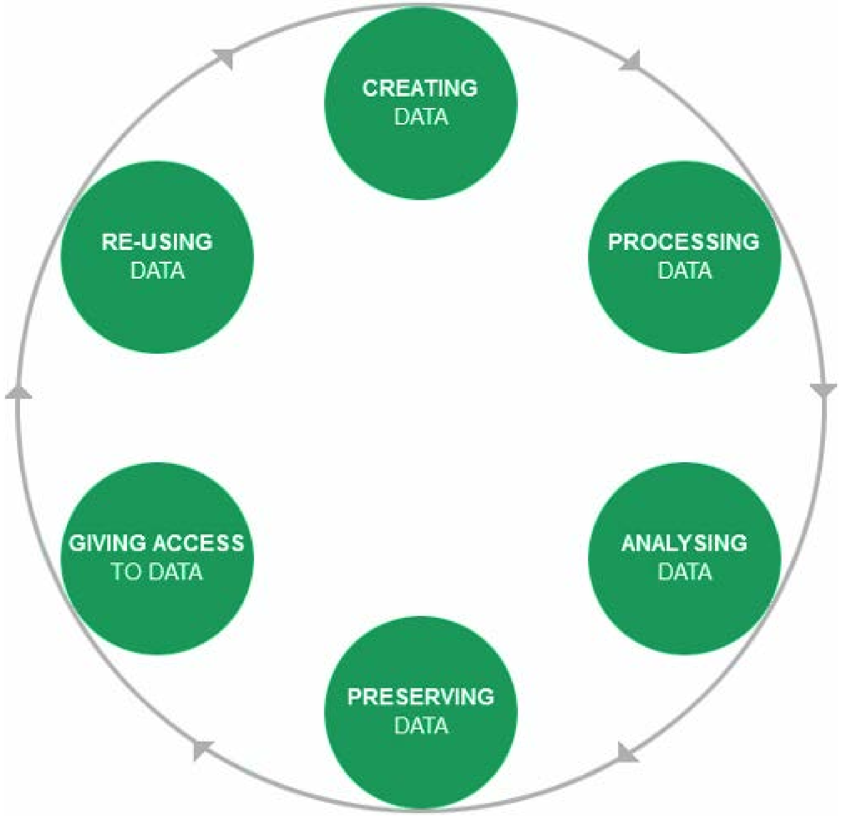
Lyon, L. (2007). *Dealing with data: Roles, rights, responsibilities, and relationships* (Consultancy Report). UK: UKOLN, University of Bath. Retrieved from http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf.

RESEARCH DATA LIFECYCLE



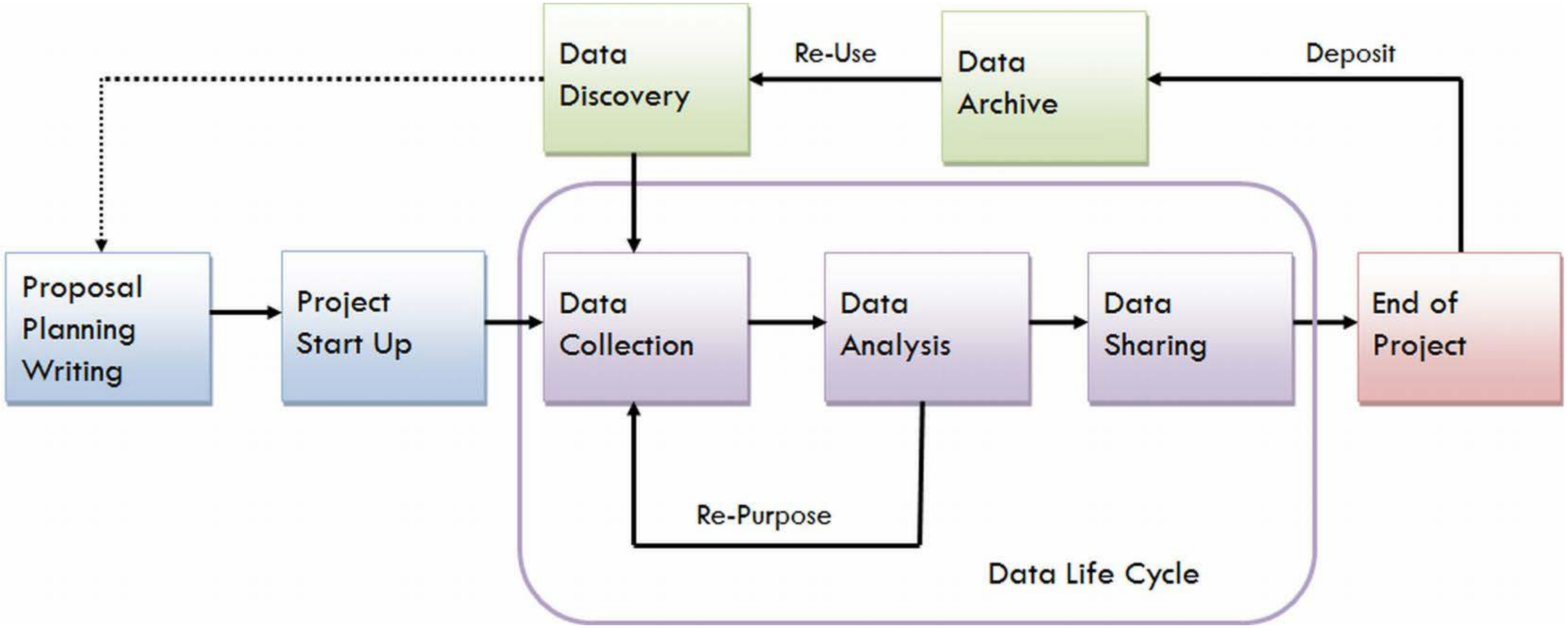
Data Documentation Initiative (DDI) Alliance. (2011, April 21). DDI Specification. Retrieved July 26, 2013, from http://www.ddialliance.org/specification

RESEARCH DATA LIFECYCLE



Corti, L. (2014). *Managing and sharing research data: A guide to good practice* (1st edition). Thousand Oaks, CA: SAGE Publications.

RESEARCH DATA LIFECYCLE

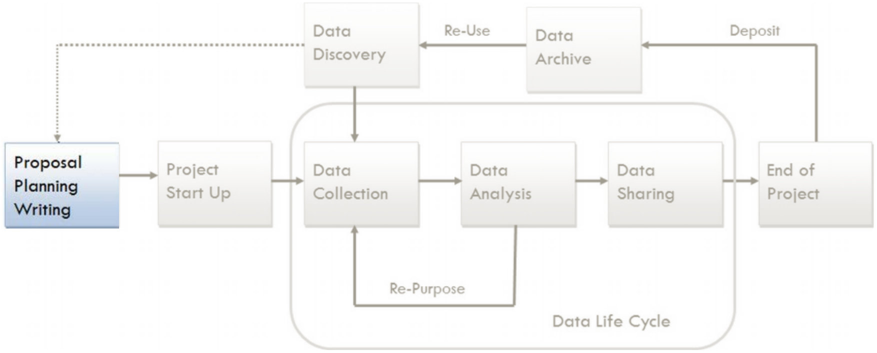


UNC THE ODUM INSTITUTE

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

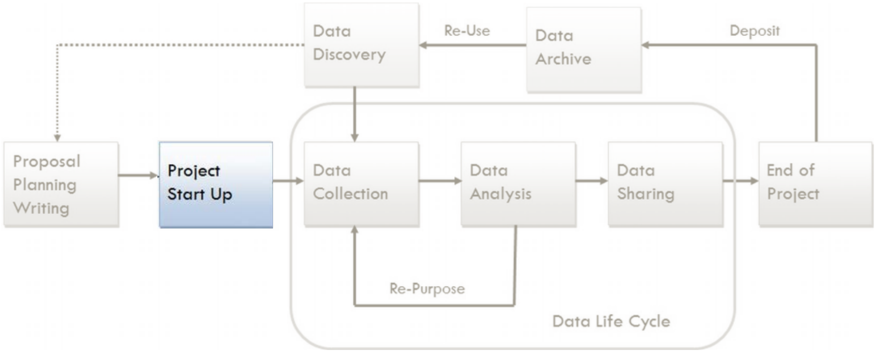
RESEARCH DATA LIFECYCLE: PROPOSAL PLANNING/WRITING

- Establish data management roles and responsibilities
- Identify expected data
- Determine required period of retention
- Consider data formats and dissemination
- Consider possible re-uses of data and how others will access them
- Select a repository that meets the needs and expected uses of the data



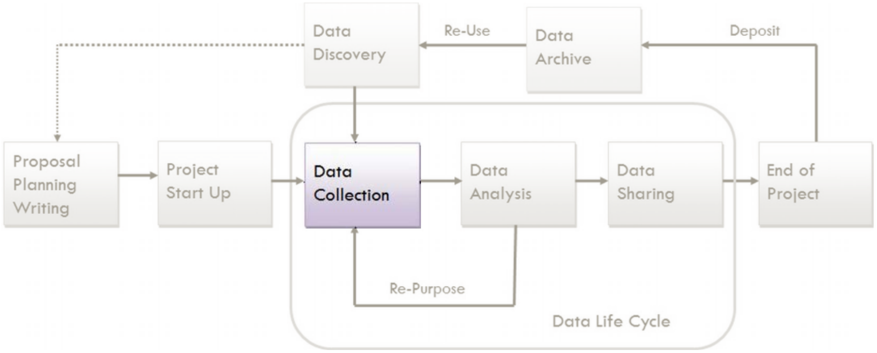
RESEARCH DATA LIFECYCLE:
PROJECT START UP

- Consider the formats and organization of data files
- Establish standard file naming conventions
- Implement measures to protect the integrity of data
- Determine the types of documentation and standard metadata that will be required to interpret and use the data
- Identify strategies to document all project decisions that affect the data



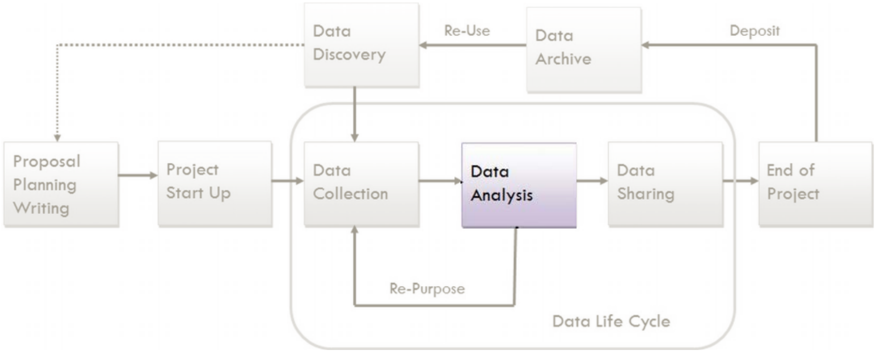
RESEARCH DATA LIFECYCLE: DATA COLLECTION

- Implement quality assurance processes to ensure dataset integrity
- Maintain accurate codebooks that contain variable and value labels
- Remain attentive to IRB requirements for the protection of human research subjects
- Implement storage and backup strategy for data



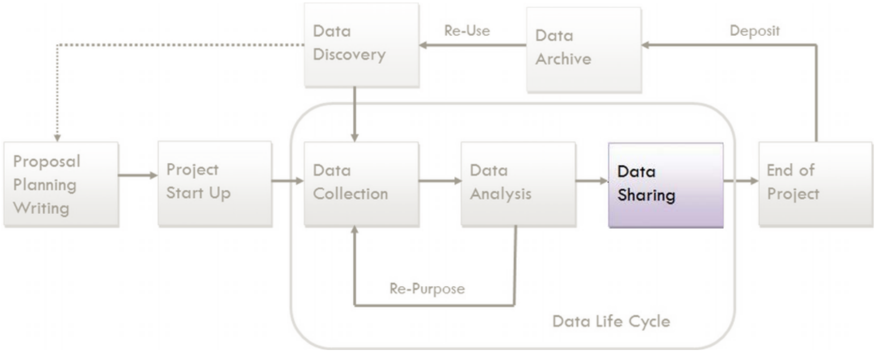
RESEARCH DATA LIFECYCLE:
DATA ANALYSIS

- Maintain explicit versions of datasets
- Use explicit version numbers in standardized file names
- Store final dataset files as read-only



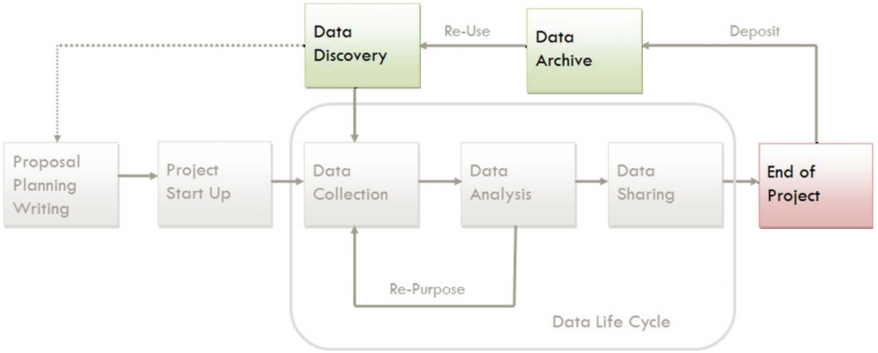
RESEARCH DATA LIFECYCLE: DATA SHARING

- Protect respondent confidentiality
- Implement strategies for limiting disclosure risk
- Restrict access and use of datasets containing protected health information (PHI) and/or personally identifiable information (PII)



RESEARCH DATA LIFECYCLE:
END OF PROJECT

- Prepare data files for submission to the data repository
- Understand the terms of use of your data in the data repository
- Cite the data!



METADATA

Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information.

National Information Standards Organization (U.S.). (2004). *Understanding metadata*. Bethesda, MD: NISO Press. Retrieved from http://www.niso.org/standards/resources/UnderstandingMetadata.pdf

WHY METADATA?

- Resource discovery
- Organizing electronic resources
- Interoperability
- Digital identification
- Archiving and preserving

National Information Standards Organization (U.S.). (2004). *Understanding metadata*. Bethesda, MD: NISO Press. Retrieved from http://www.niso.org/standards/resources/UnderstandingMetadata.pdf

METADATA SCHEMES & ELEMENT SETS

MODS

CCO Commons
Cataloging Cultural Objects

TEI
Text Encoding Initiative

DataCite

DDI®

PREMIS
PRESERVATION METADATA MAINTENANCE ACTIVITY

METS
Metadata Encoding & Transmission Standard
Official Web Site

Dublin Core® Metadata Initiative
Making it easier to find information.

DATA QUALITY

The ***replication standard*** holds that 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** (p. 444).

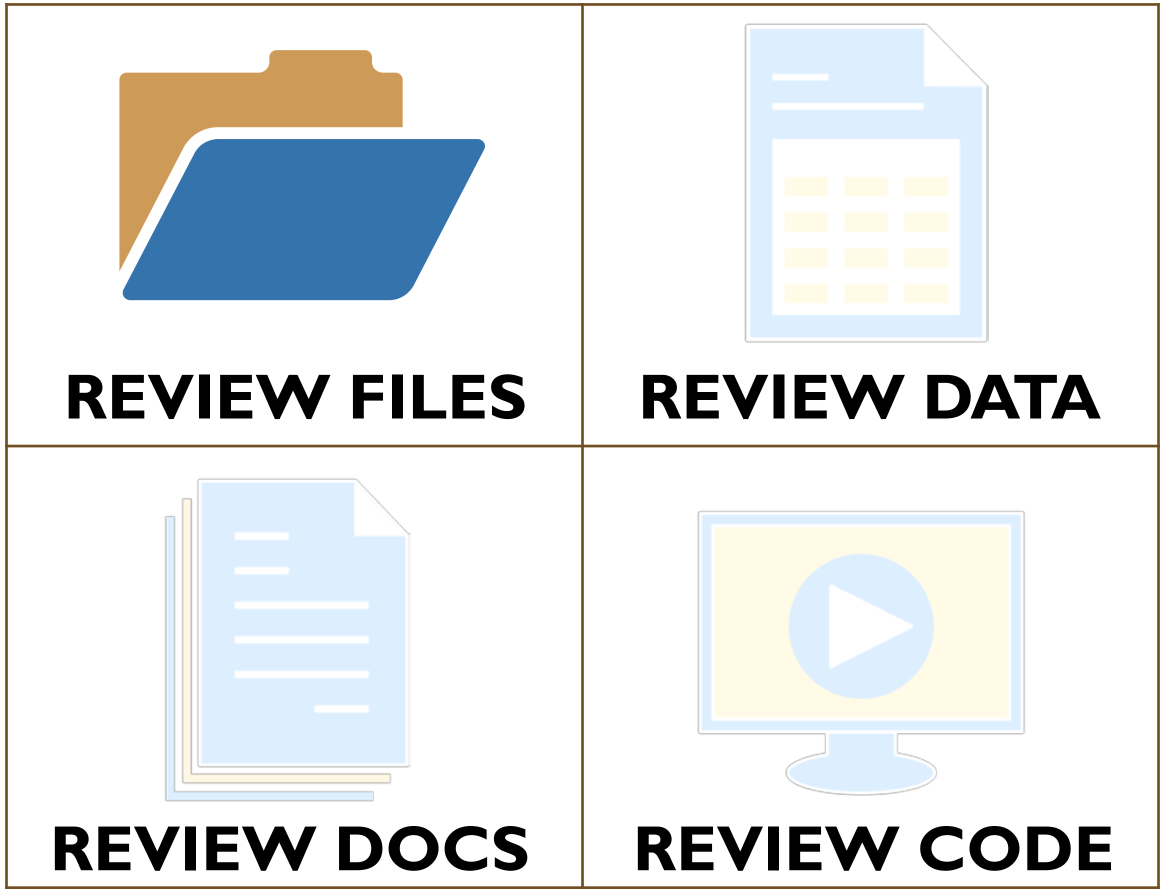
King, G. (1995). Replication, replication. *PS: Political Science & Politics*, *28*(3), 444–452. http://doi.org/10.2307/420301

DATA QUALITY

REVIEW FILES	REVIEW DATA
REVIEW DOCS	REVIEW CODE

Peer, L., Green, A., & Stephenson, E. (2014). Committing to data quality review. *International Journal of Digital Curation*, *9*(1). http://doi.org/10.2218/ijdc.v9i1.317

DATA QUALITY



REVIEW FILES

REVIEW DATA

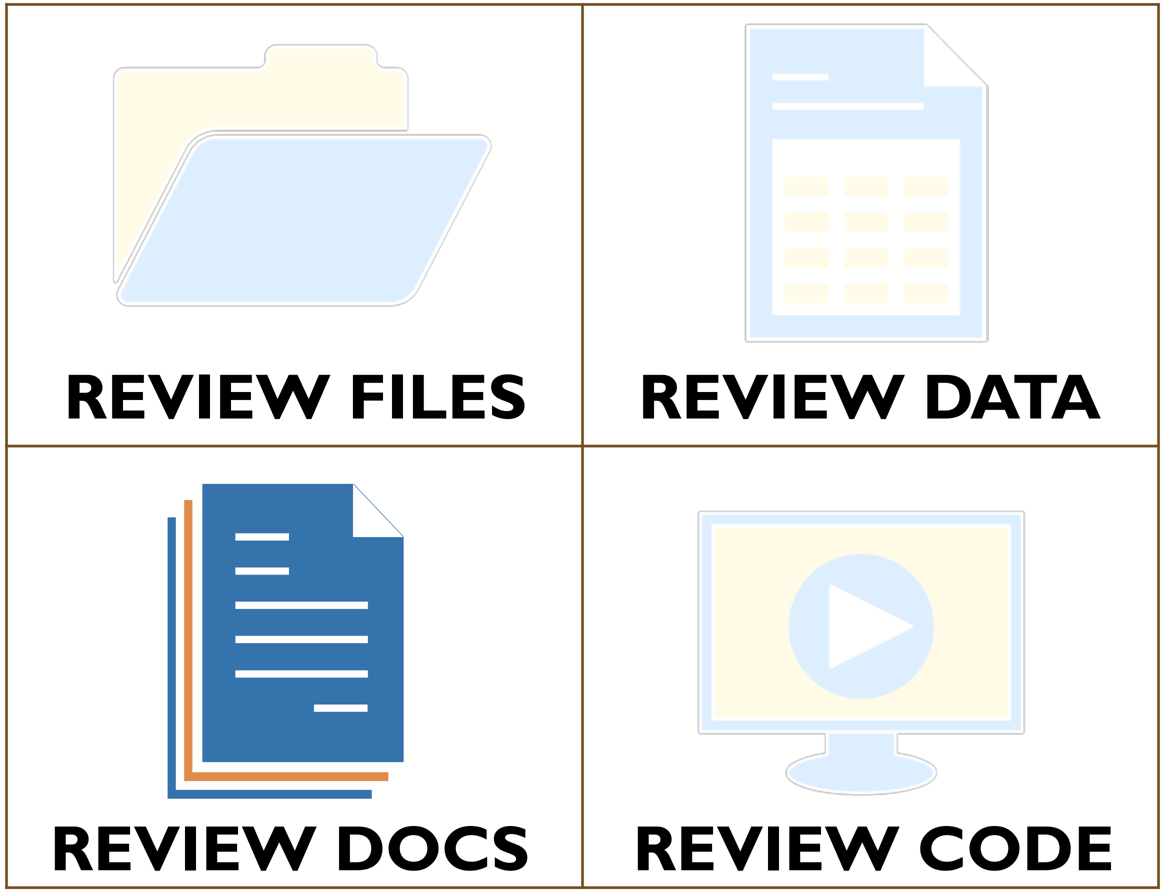
REVIEW DOCS

REVIEW CODE

- ✓ Assign persistent IDs
- ✓ Create study citation
- ✓ Record file details
- ✓ Check that all files are present
- ✓ Verify file content and format matches
- ✓ Create preservation copies
- ✓ Implement migration strategy
- ✓ Monitor bits

Peer, L., Green, A., & Stephenson, E. (2014). Committing to data quality review. *International Journal of Digital Curation*, *9*(1). http://doi.org/10.2218/ijdc.v9i1.317

DATA QUALITY



REVIEW FILES

REVIEW DATA

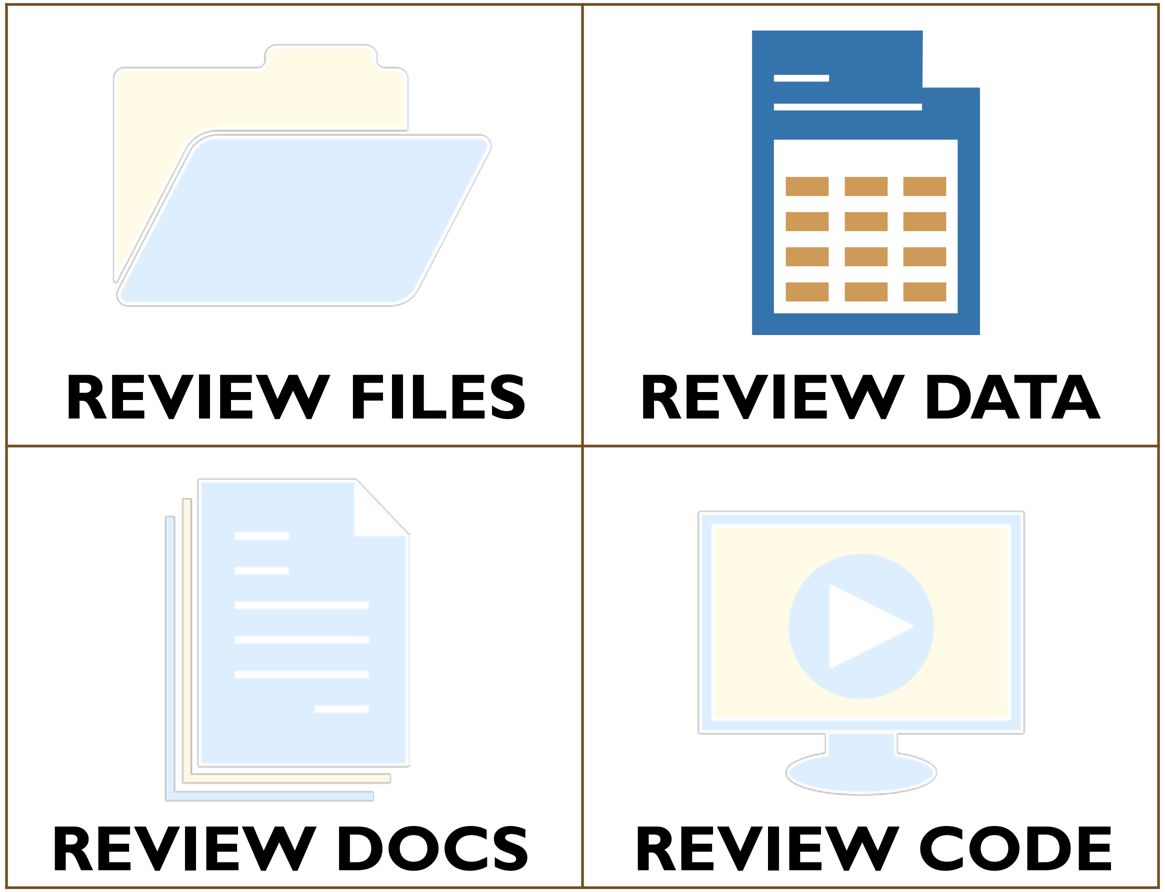
REVIEW DOCS

REVIEW CODE

- ✓ Confirm comprehensive descriptive information for informed reuse including methodology and sampling information
- ✓ Link to other research products

Peer, L., Green, A., & Stephenson, E. (2014). Committing to data quality review. *International Journal of Digital Curation*, *9*(1). http://doi.org/10.2218/ijdc.v9i1.317

DATA QUALITY



REVIEW FILES

REVIEW DATA

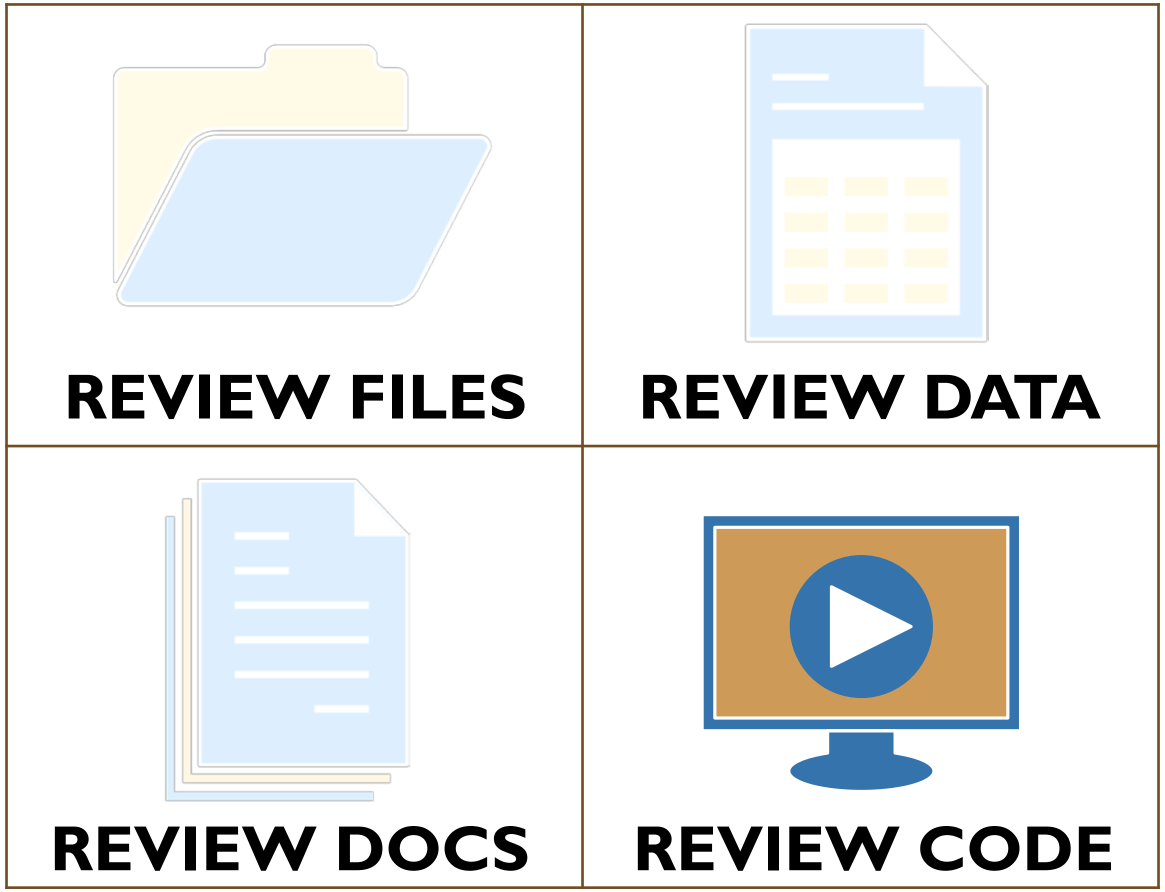
REVIEW DOCS

REVIEW CODE

- ✓ Check for undocumented variable and value information or out of range codes
- ✓ Review data for confidentiality issues

Peer, L., Green, A., & Stephenson, E. (2014). Committing to data quality review. *International Journal of Digital Curation*, *9*(1). http://doi.org/10.2218/ijdc.v9i1.317

DATA QUALITY



REVIEW FILES

REVIEW DATA

REVIEW DOCS

REVIEW CODE

- ✓ Check and verify code for data analysis and replication

Peer, L., Green, A., & Stephenson, E. (2014). Committing to data quality review. *International Journal of Digital Curation*, *9*(1). http://doi.org/10.2218/ijdc.v9i1.317

DATA MANAGEMENT PLANNING

Developing a data management plan (DMP) enables you to achieve the benefits of managing and sharing data, which include…

- Being able to find and understand your data
- Avoiding unnecessary duplication of data collection efforts
- Maintaining data underlying published results, allowing for validation and replication
- Increasing the visibility and impact of your data
- Promoting new research and collaborations

DMP CHECKLIST:
ADMINISTRATIVE INFORMATION

☑ ID

☑ Funder + Grant reference number

☑ Project name

☑ Project description

☑ Principal investigator(s)

☑ DMP date

DCC. (2013). *Checklist for a Data Management Plan.* v.4.0. Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/data-management-plans

DMP CHECKLIST:
DATA COLLECTION

- ☑ What data will you collect or create?
- ☑ How will the data be collected or created?

DCC. (2013). *Checklist for a Data Management Plan.* v.4.0. Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/data-management-plans

DMP CHECKLIST:
DOCUMENTATION & METADATA

☑ What documentation and metadata will accompany the data?

DCC. (2013). *Checklist for a Data Management Plan.* v.4.0. Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/data-management-plans

DMP CHECKLIST:
ETHICS & LEGAL COMPLIANCE

☑ How will you manage any ethical issues?

☑ How will you manage copyright and Intellectual Property Rights issues?

DCC. (2013). *Checklist for a Data Management Plan.* v.4.0. Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/data-management-plans

DMP CHECKLIST:
STORAGE & BACKUP

☑ How will the data be stored and backed up during the research?

☑ How will you manage access and security?

DCC. (2013). *Checklist for a Data Management Plan.* v.4.0. Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/data-management-plans

DMP CHECKLIST:
SELECTION & PRESERVATION

- ☑ Which data should be retained, shared, and/or preserved?
- ☑ What is the long-term preservation plan for the dataset?

DCC. (2013). *Checklist for a Data Management Plan.* v.4.0. Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/data-management-plans

DMP CHECKLIST:
DATA SHARING

- ☑ How will you share the data?
- ☑ Are any restrictions on data sharing required?

DCC. (2013). *Checklist for a Data Management Plan.* v.4.0. Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/data-management-plans

DMP CHECKLIST:
RESPONSIBILITIES & RESOURCES

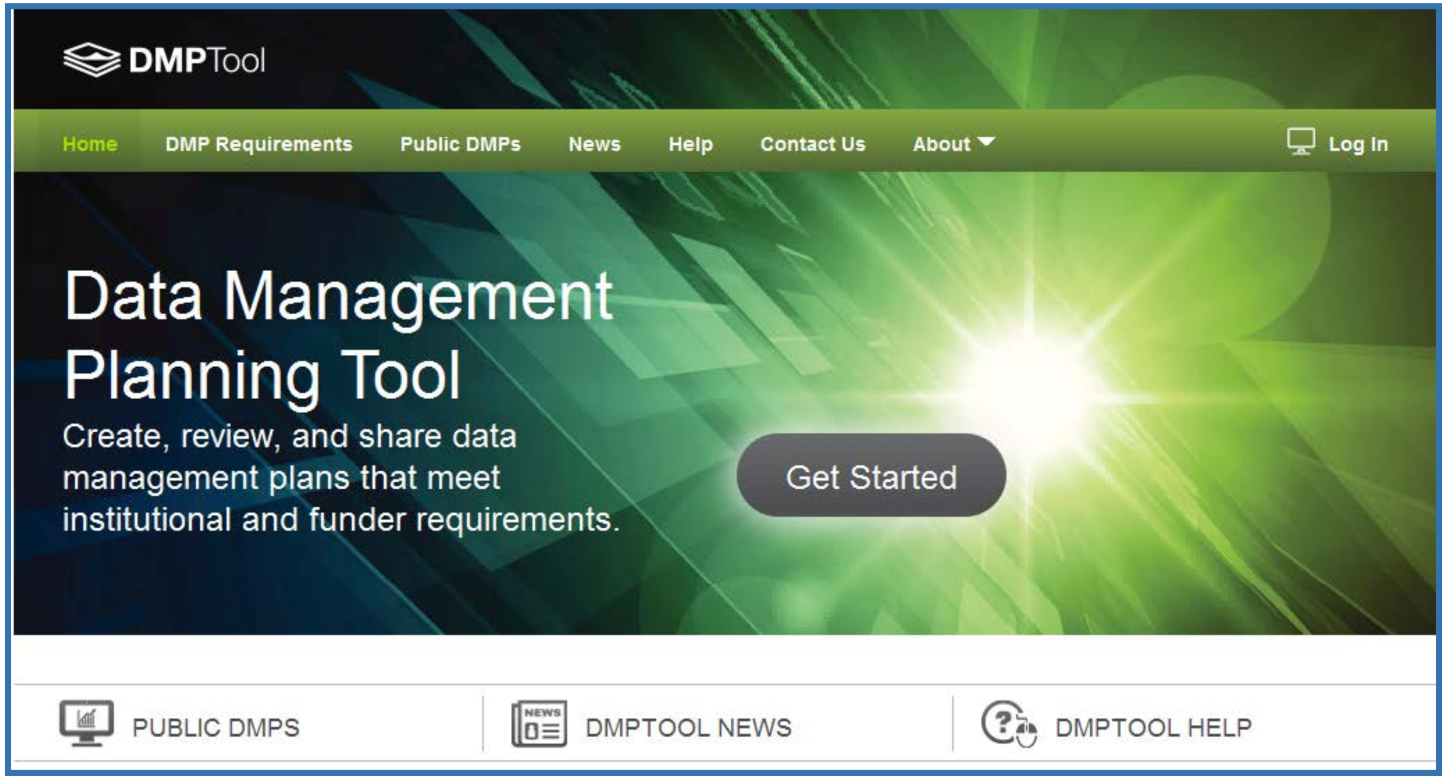
☑ Who will be responsible for data management?

☑ What resources will you require to deliver your plan?

DCC. (2013). *Checklist for a Data Management Plan.* v.4.0. Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/data-management-plans

DMPTOOL

https://dmptool.org



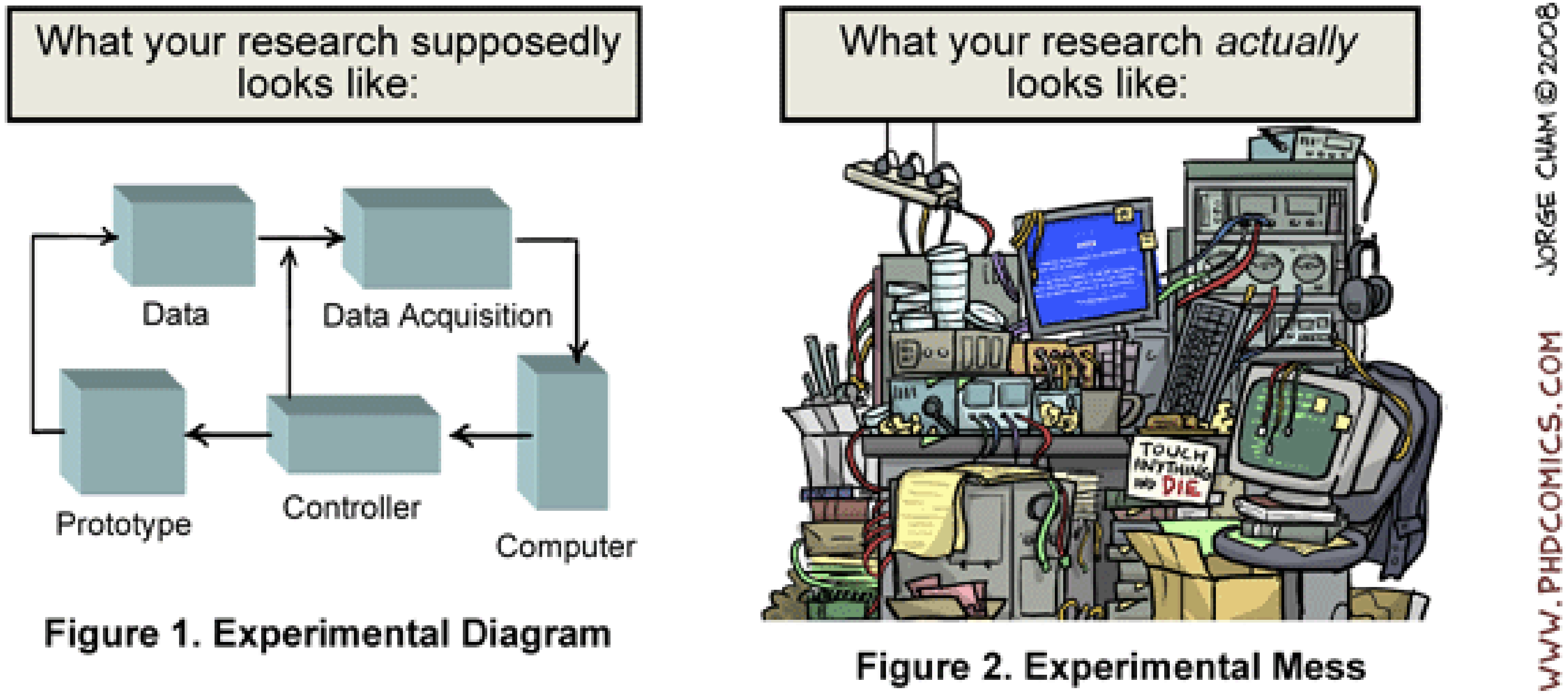
Data Management Planning Tool

Create, review, and share data management plans that meet institutional and funder requirements.

DATA SHARING OBSTACLES

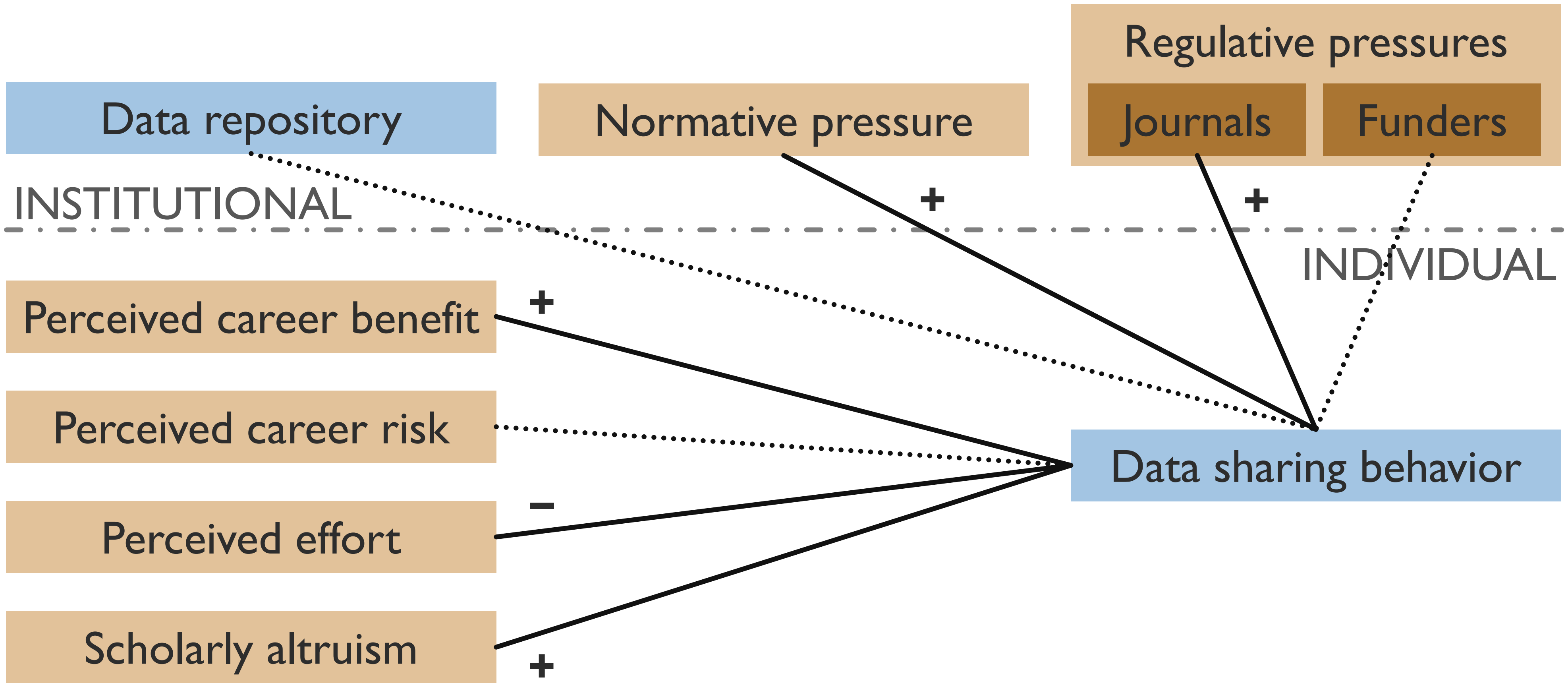
- Time and effort to make data shareable
- Perceived risks from loss of control of the data
- Data contain sensitive information
- Data ownership may be unclear or problematic
- Lack of incentives for sharing data

DATA SHARING OBSTACLES


What your research supposedly looks like:
Data
Data Acquisition
Prototype
Controller
Computer
Figure 1. Experimental Diagram
What your research *actually* looks like:
TOUCH ANYTHING AND DIE
Figure 2. Experimental Mess
JORGE CHAM ©2008
WWW.PHDCOMICS.COM

Cham, J. (2008). Research diagram/research reality. *Piled Higher and Deeper.* Retrieved from http://www.phdcomics.com/comics/archive.php?comicid=961

DATA SHARING FACTORS



Kim, Y., & Stanton, J. M. (2015). Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology.* http://doi.org/10.1002/asi.23424

THE TRUSTED REPOSITORY

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

THE TRUSTED REPOSITORY

- Accept responsibility for the long-term maintenance of resources
- Have an organizational system that supports both the long-term viability of the repository and its contents
- Demonstrate fiscal responsibility and sustainability
- Design its system in accordance with commonly accepted conventions and standards

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

THE TRUSTED REPOSITORY

- ➡ Establish methodologies for system evaluation that meet community expectations for trustworthiness
- ➡ Be depended on to carry out its long-term responsibilities to depositors and users
- ➡ Have policies, practices, and performance that can be audited and measured

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

STANDARDS OF TRUST

Data Seal of Approval (DSA)

Data Asset Framework

DIN 31644: Criteria for Trustworthy Digital Archives

DRAMBORA

ISO 16363: Audit and Certification of Trustworthy Digital Repositories

The nestor Seal for Trustworthy Digital Archives

Repository Audit and Certification DSA-WDS Partnership Working Group

DATA SEAL OF APPROVAL: CRITERIA

☑ The research data can be found on the Internet

☑ The research data are accessible, while taking into account relevant legislation with regard to personal information and intellectual property of the data

☑ The research data are available in a usable format

☑ The research data are reliable

☑ The research data are citable

2014
DSA
2015

UNC
THE ODUM INSTITUTE

DATA SEAL OF APPROVAL: STATEMENTS OF COMPLIANCE

0 = N/A	Not applicable
1 = No	We have not considered this yet
2 = Theoretical	We have a theoretical concept
3 = In Progress	We are in the implementation phase
4 = Implemented	This guideline has been fully implemented for the needs of our repository

Data Producer

Data Repository

Data Consumer

DATA SEAL OF APPROVAL: GUIDELINES

Data Producer

1. The data producer deposits the data in a repository with sufficient information to assess the quality of the data and compliance with disciplinary and ethical norms. (3)
2. The data producer provides the data in formats recommended by the data repository. (3)
3. The data producer provides the data together with the metadata requested by the data repository. (4)

DATA SEAL OF APPROVAL: GUIDELINES

Data Repository

4. The data repository has an explicit mission in the area of digital archiving and promulgates it. (4)
5. The data repository uses due diligence to ensure compliance with legal regulations and contracts including, when applicable, regulations governing the protection of human subjects. (4)
6. The data repository applies documented processes and procedures for managing data storage. (4)

DATA SEAL OF APPROVAL: GUIDELINES

Data Repository

7. The data repository has a plan for long-term preservation of its digital assets. (3)
8. Archiving takes place according to explicit work flows across the data life cycle. (3)
9. The data repository assumes responsibility from the data producers for access and availability of digital objects. (4)

DATA SEAL OF APPROVAL: GUIDELINES

Data Repository

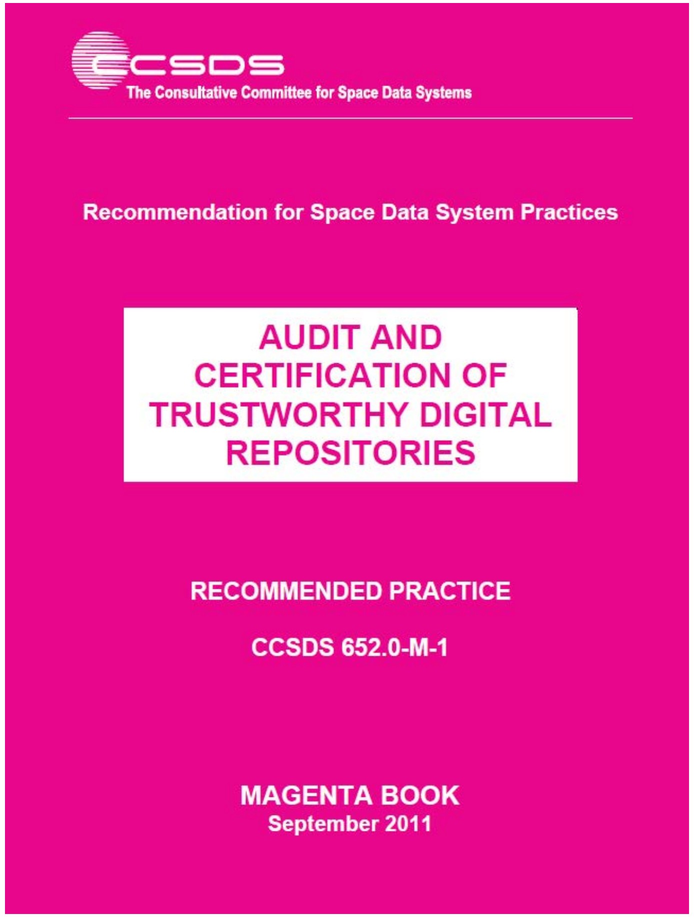
10. The data repository enables users to utilize the data and refer to them in a persistent way. (3)
11. The data repository ensures the integrity of the digital objects and metadata. (3)
12. The data repository ensures the authenticity of the digital objects and the metadata. (3)
13. The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS. (3)

DATA SEAL OF APPROVAL: GUIDELINES

Data Consumer

14. The data consumer complies with access regulations set by the data repository. (4)
15. The data consumer conforms to and agrees with any codes of conduct that are generally accepted in the relevant sector for the exchange and proper use of knowledge and information. (4)
16. The data consumer respects the applicable licenses of the data repository regarding the use of the data. (4)

ISO 16363: AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES



CCSDS

The Consultative Committee for Space Data Systems

Recommendation for Space Data System Practices

AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES

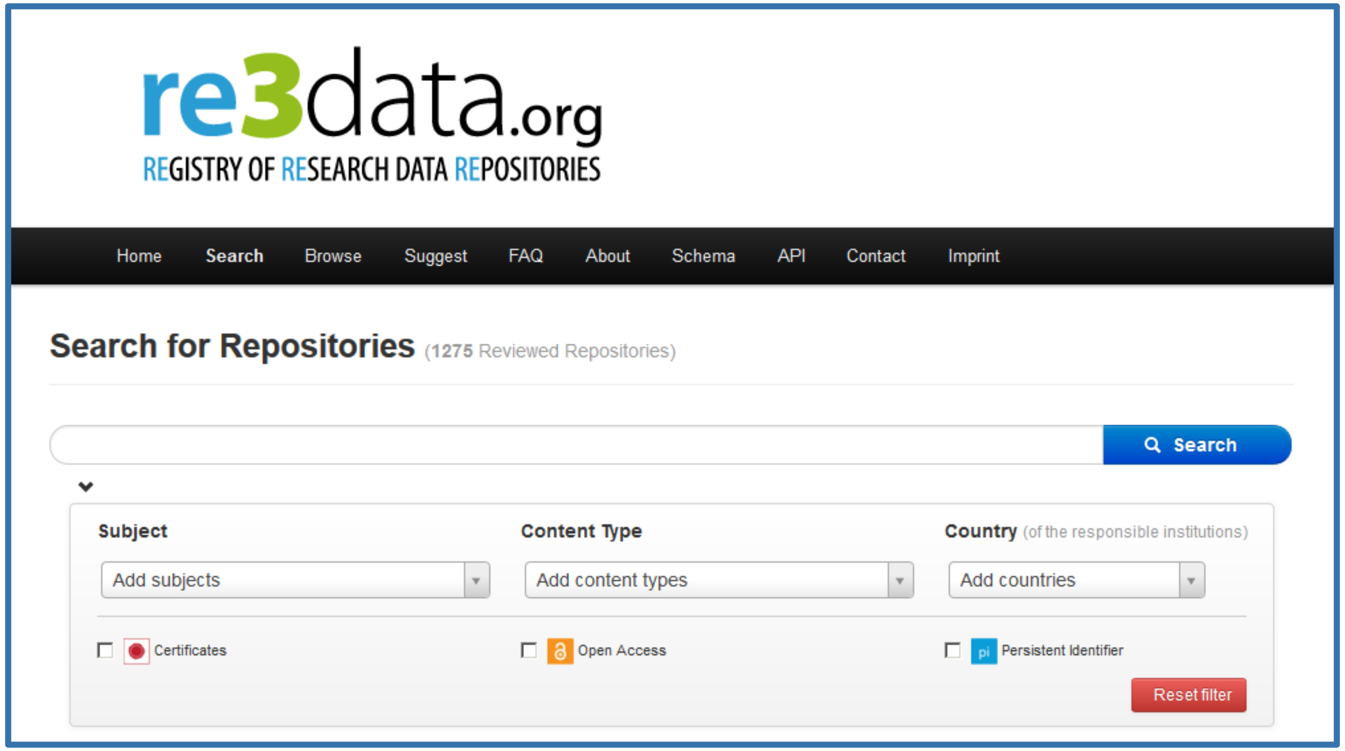
RECOMMENDED PRACTICE

CCSDS 652.0-M-1

MAGENTA BOOK
September 2011

FINDING A TRUSTWORTHY REPOSITORY

http://www.re3data.org/



re3data.org

REGISTRY OF RESEARCH DATA REPOSITORIES

Search for Repositories (1275 Reviewed Repositories)

Subject: Add subjects

Content Type: Add content types

Country (of the responsible institutions): Add countries

□ Certificates

□ Open Access

□ Persistent Identifier

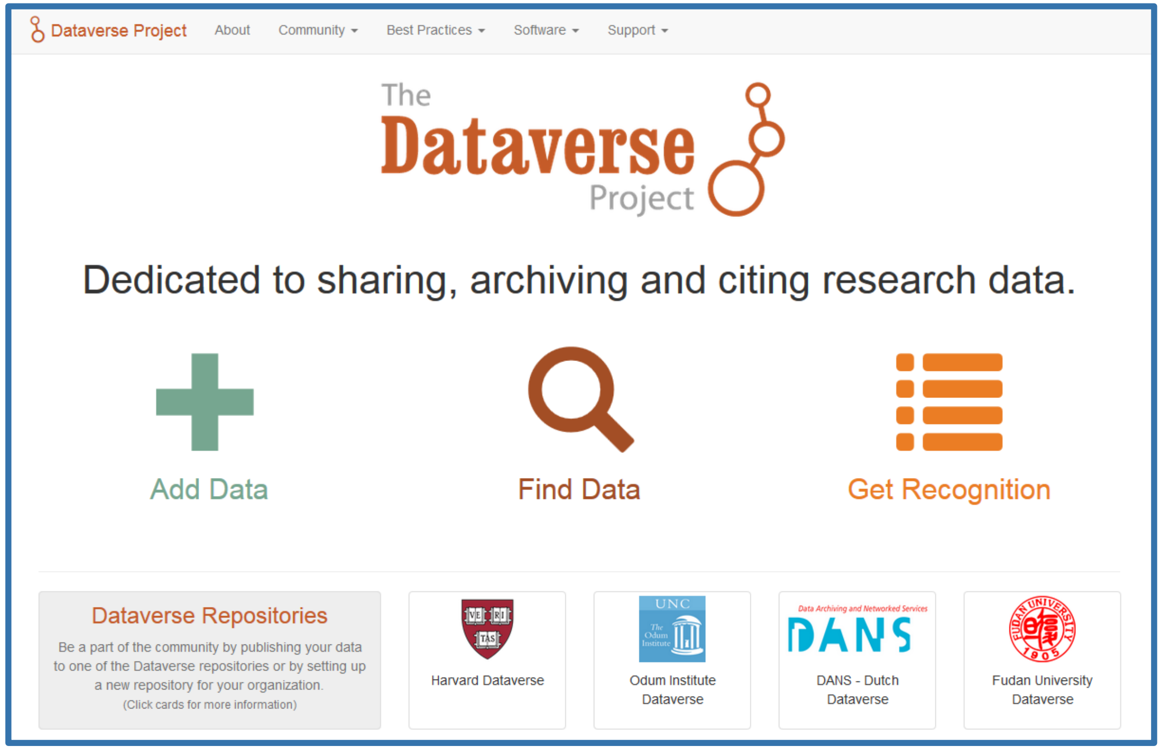
FINDING A TRUSTWORTHY REPOSITORY

Using re3data.org, identify a trustworthy data 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?
6. *What other criteria are important for your data?*

THE DATAVERSE PROJECT

http://dataverse.org



Dedicated to sharing, archiving and citing research data.

Dataverse Repositories

Be a part of the community by publishing your data to one of the Dataverse repositories or by setting up a new repository for your organization.
(Click cards for more information)

THE DATAVERSE PROJECT

- Developed at Harvard University’s Institute for Quantitative Social Science (IQSS)
- Open source web application for publishing, citing, analyzing, and preserving research data

The
Dataverse
Project

THE DATAVERSE PROJECT

- 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
- Linked with the Open Journal Systems publishing platform

The Dataverse Project

OTHER CURATION TOOLS

Open Science Framework ▪ https://osf.io
Single online portal for research project management with add-ons to connect to external tools for storage, security, and citation

Archivematica ▪ https://www.archivematica.org
Software tool that supports data processing for archival ingest in compliance with ISO-OAIS functional model

DataTags (beta) ▪ http://datatags.org
Prototype tool for supporting compliance with laws and contractual agreements that govern sensitive data sharing

OTHER CURATION TOOLS

CDART ▪ https://www2.cscc.unc.edu/home/cdart

Carolina Data Acquisition and Reporting Tool. Research data management system for clinical research data that supports clinical trials, patient registries, and observation studies

WC3 PROV ▪ http://www.w3.org/TR/prov-overview/

Data model for interoperable exchange of provenance information

Data Carpentry ▪ http://datacarpentry.org

Spin-off of Software Carpentry for teaching basic concepts, skills, and tools for working with data

QUESTIONS?

Thu-Mai Christian
tlchristian@unc.edu

Sophia Lafferty-Hess
slaffer@email.unc.edu

www.odum.unc.edu

@Odum_Institute

OdumInstitute