Data Sharing & FAIRness

Jennifer Moore Head of Data Services j.moore@wustl.edu

https://library.wustl.edu/services/data/

Why you have to care…
…or will soon

Data management & sharing plans are becoming mandatory

- Funders
- Journals
- Office of Science & Technology Policy mandate
- Community expectations

Data sharing and public trust in research

Public trust in research findings increases with data sharing – survey by Pew Research Center

3. Americans say open access to data and independent review inspire more trust in research findings

The Pew Research Center survey asked about several factors that could potentially increase – or decrease – trust in research findings and recommendations. The two steps that inspire the most confidence among members of the public are open access to data and an independent review.

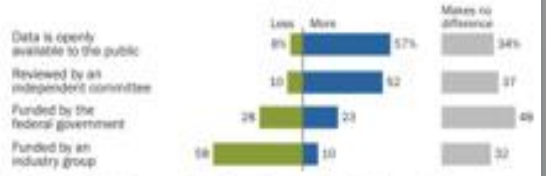
A majority of U.S. adults (57%) say they trust scientific research findings *more* if the researchers make their data publicly available. Another 34% say that makes no difference, and just 8% say they are less apt to trust research findings if the data is released publicly.

About half the public (52%) say they trust scientific findings more if the findings have been reviewed by an independent committee.

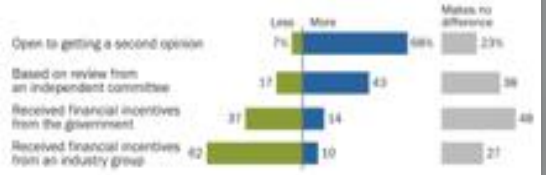
Industry funding stands out as a factor Americans say leads to lower trust. A majority of Americans (58%) say they trust scientific findings less if they know the research was funded by industry groups.

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 ...

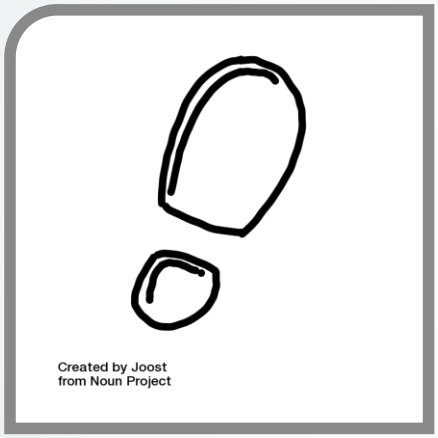


% of U.S. adults who say when they hear each of the following, they trust a science practitioner's recommendation ...



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



Advancing Research & Your Legacy

- Shared research data advances scientific discovery
- Preserve all the work you've done and get cited!
- Gets citations

Exercise on your own

Look for a funder or publisher that is relevant to your work

Review their data management and/or sharing requirements

What questions do you have?

Making shared data FAIR

Findable – Is in an indexed repository, with a unique, persistent ID, and has rich metadata

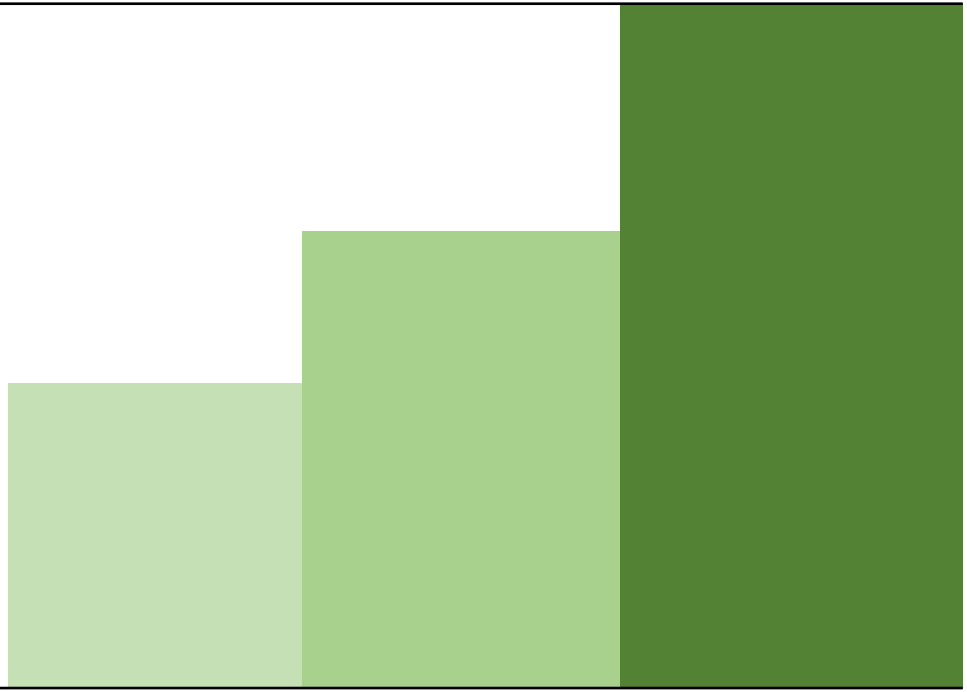
Accessible – repo uses open, standard protocols so the metadata and data can be accessed

Interoperable – data are in formal, standard, open application languages

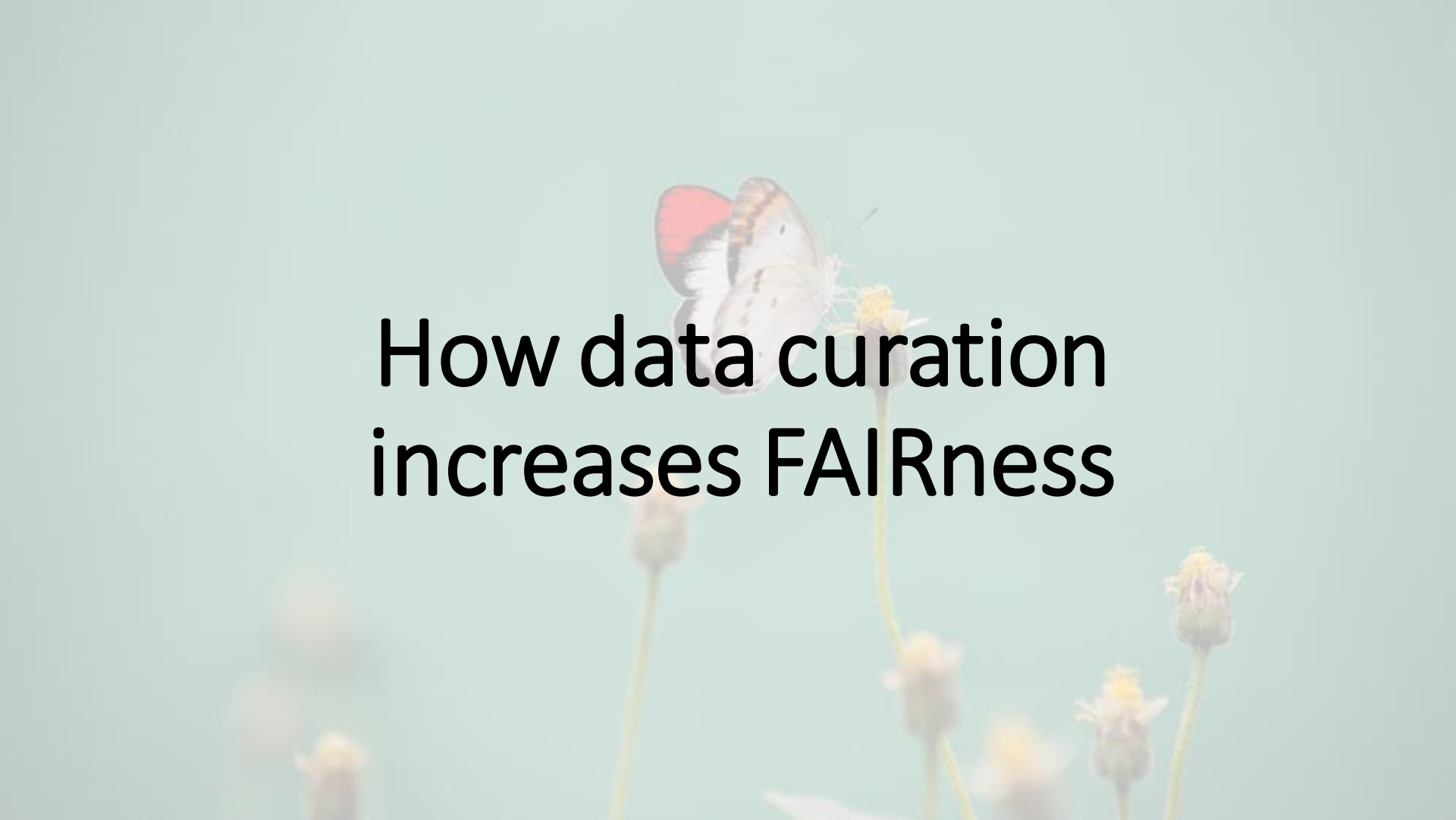
Reusable – well documented, explicit provenance, open licenses, follows community standards

Ways to increase FAIRness

TREATMENTS	GOOD	BETTER	BEST
Transformed to an open format			
Domain-specific, enhanced metadata			
Methods documentation			
Checked by a curator			
Persistent identifier			
Indexed repository			
Backed up			
Monitored for change			
Basic descriptive metadata			



How data curation increases FAIRness



Value of Curation

- Longevity
- Accessibility
- Transparency
- Reusability
- Advancement
- Availability
- Accountability
- Citability

Increase FAIRness

Findable | Accessible | Interoperable | Reusable

What do we mean by data curation?

Data Curation
facilitates sharing & preservation of data created in within a research project through:

Selection
Management
Documentation
Identification
Transformation
Archiving

What can data can be curated?

IDEAL	NOT IDEAL
At a milestone or end of project	In active collection
Owned by the researcher	Has rights restrictions
Open	Proprietary
Well documented	Undocumented
Ethical	People at risk

Increase FAIRness

Findable | Accessible | Interoperable | Reusable

Steps you can take to prepare to share

Documentation of methods

Readme file

Codebook

De-identify

Resources for data curation

Data Services at WashU:
https://library.wustl.edu/data/

Data Curation Network:
https://datacurationnetwork.org/

Deidentification

Ethics of Identity

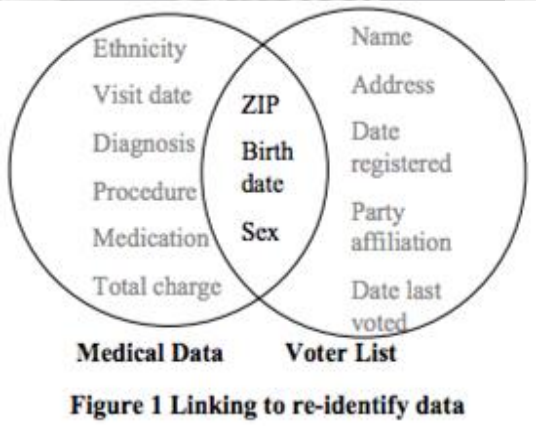


Figure 1 Linking to re-identify data

ALWAYS adhere to institutional review board (**IRB**)/ethics committee guidelines and recommendations and professional standards

k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY[1]

LATANYA SWEENEY

ool of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
E-mail: latanya@cs.cmu.edu

Received May 2002

Consider a data holder, such as a hospital or a bank, that has a privately held collection of person-specific, field structured data. Suppose the data holder wants to share a version of the data with researchers. How can a data holder release a version of its private data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful? The solution provided in this paper includes a formal protection model named *k*-anonymity and a set of accompanying policies for deployment. A release provides *k*-anonymity protection if the information for each person contained in the release cannot be distinguished from at least *k*-1 individuals whose information also appears in the release. This paper also examines re-identification attacks that can be realized on releases that adhere to *k*-anonymity unless accompanying policies are respected. The *k*-anonymity protection model is important because it forms the basis on which the real-world systems known as Datafly, µ-Argus and *k*-Similar provide guarantees of privacy protection.

Keywords: data anonymity, data privacy, re-identification, data fusion, privacy.

HIPPA – 18 Identifiers

- Name
- Address
- All elements of dates related to an individual
- Telephone numbers
- Fax number
- Email address
- Social Security Number
- Medical record number
- Health plan beneficiary number
- Account number
- Certificate or license number
- Any vehicle or other device serial number
- Web URL
- Internet Protocol (IP) Address
- Finger or voice print
- Photographic
- Any other unique characteristic

Exercise 1: on your own

REMOVE ALL IDENTIFIERS

NOTE THE STEPS YOU TOOK

Tools which manage identity

RedCap redcap.wustl.edu

Easily tag fields as identifiers – use for names, dates, and other PHI

Seamlessly remove all tagged identifier fields at data export

Can simply hide these tagged identifier fields from collaborators that do not need to see them (e.g. statistician)

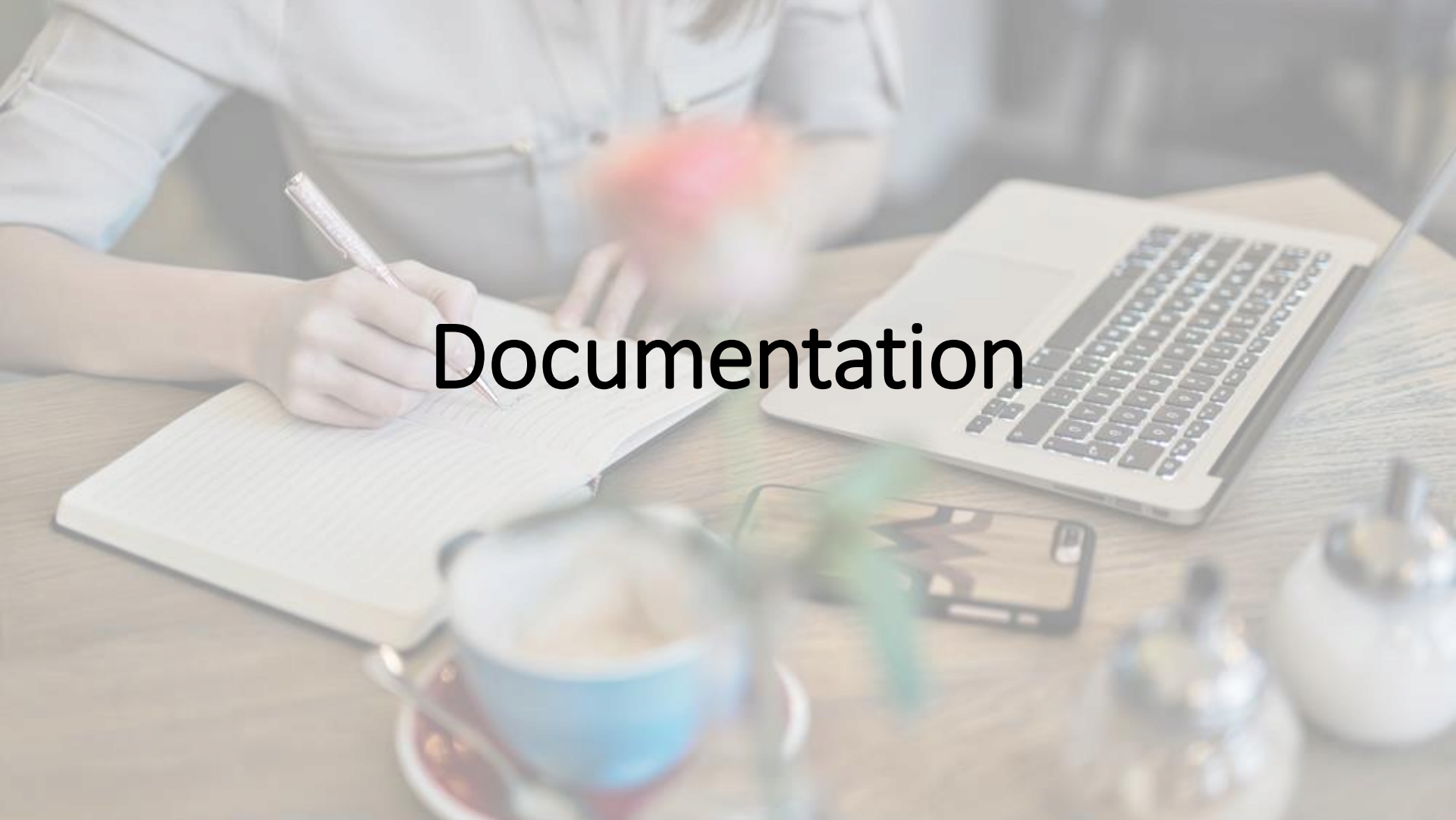
Resources to understand subject protection

Office of the Vice Chancellor for Research (OVCR)

WashU Institutional Review Board (IRB)

Health Insurance Portability and Accountability Act (HIPPA)

Family Educational Rights Act (FERPA)


Documentation

Data Dictionary/Metadata

Create a README to include with your data package

Typically includes:

Variable name	Variable meaning	Variable units	Variable format	Variable values & meanings	Known issues	Relationship to other variables	Null value indicator	Wx5

Exercise 2: on your own

DOWNLOAD README TEMPLATE

FILL IN AS MUCH AS YOU CAN BASED ON YOUR PROJECT DATA

Resources for documentation

Digital Research Materials Repository – find a readme template text file

Open Science Framework (OSF) – collect, document, collaborate Free, browser-based

LabArchives – collect, document, collaborate - integrates well with Office – WashU Subscription, browser-based

Jupyter Notebooks – document and collaborate, works well with R and and Python Free and open source

Collectica – document - works well with Excel

Sharing

in a Repository

Finding the right repository

- Advantages of domain repository
- Often supported in that field
- Some publishers prefer
- Specific metadata
- Community driven
- Often costs are associated

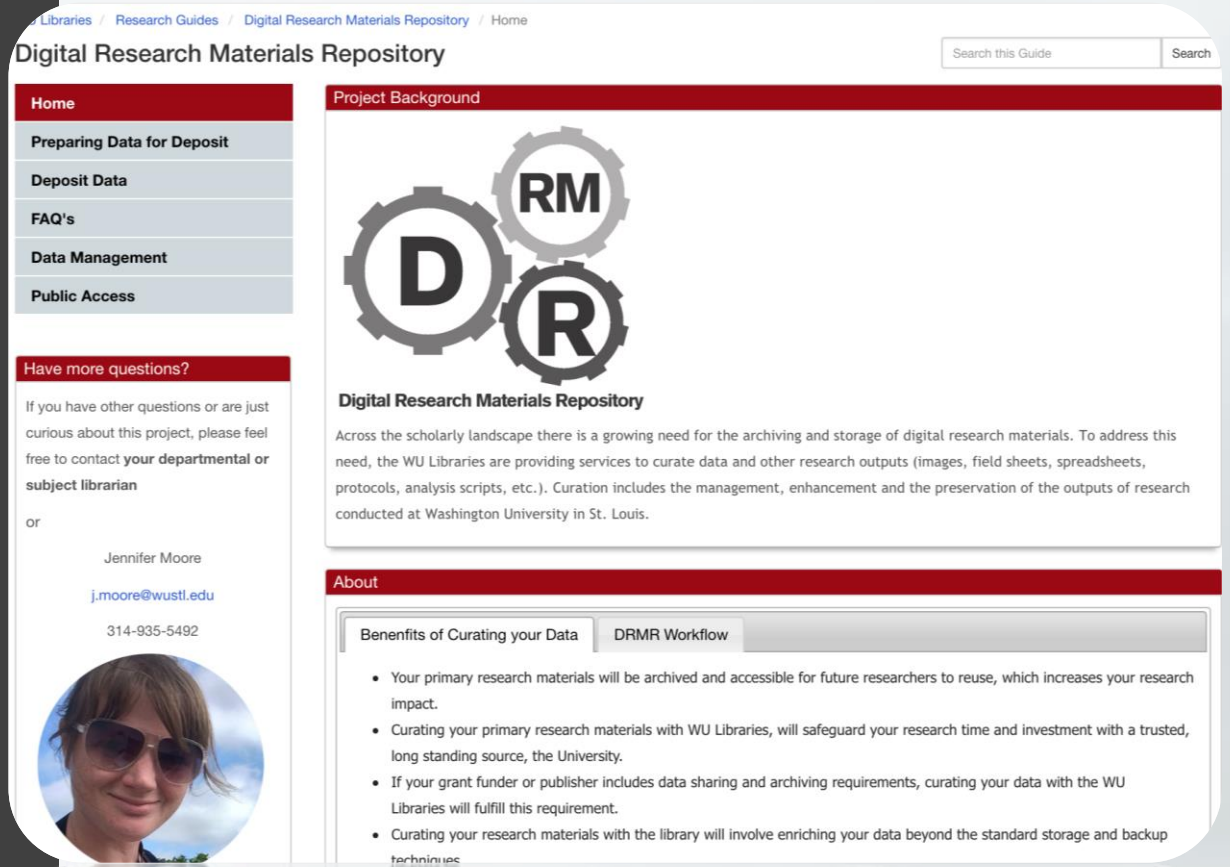
Finding the right repository

- Advantages of a generalized repository
- Reaches a wider audience
- Many research institutions have repositories
- Institutional repositories often don’t have fees
- Institutional repositories are often committed to long-term preservation and have curation services

What	You	Curator	Repository
Indexed			x
Persistent identifier			x
Backed up			x
Monitored for change			x
Descriptive metadata	x	x	x
Methods documentation	x		
Reviewed to meet FAIR principles		x	
Transformed to an open format	x	x	
Domain-specific, enhance metadata	x	x	
Distributed Copies	x	x	x

WashU offers:

- Local curation services
- Networked curation services
- Persistent ID
- Backed-up
- Monitored for change
- Descriptive metadata
- Support for domain specific metadata
- Format transformation
- Distributed copies



https://libguides.wustl.edu/drmr

Exercise 3: look at repositories

SEARCH FOR RE3DATA

GO TO YOUR DOMAIN AND REVIEW POTENTIAL REPOSITORIES

Tools which manage identity

Re3Data

Re3data.org is a global registry of research data repositories that covers research data repositories from different academic disciplines. It presents repositories for the permanent storage and access of data sets to researchers, funding bodies, publishers and scholarly institutions.

Data management contacts

Danforth Campus

- Jennifer Moore – j.moore@wustl.edu
- Dorris Scott – d.scott@wustl.edu

Medical Campus

- Christopher Sorensen sorensenc@wustl.edu
- Marcy Vana - vanam@wustl.edu

j.moore@wustl.edu