Research Project Workshop 02:

Research Data Management (RDM) Overview

March15, 2017

Instructor: Letisha Wyatt, PhD

tinyurl.com/z8qn8o4

agenda & learning outcomes

1 Understand the elements of the data life cycle

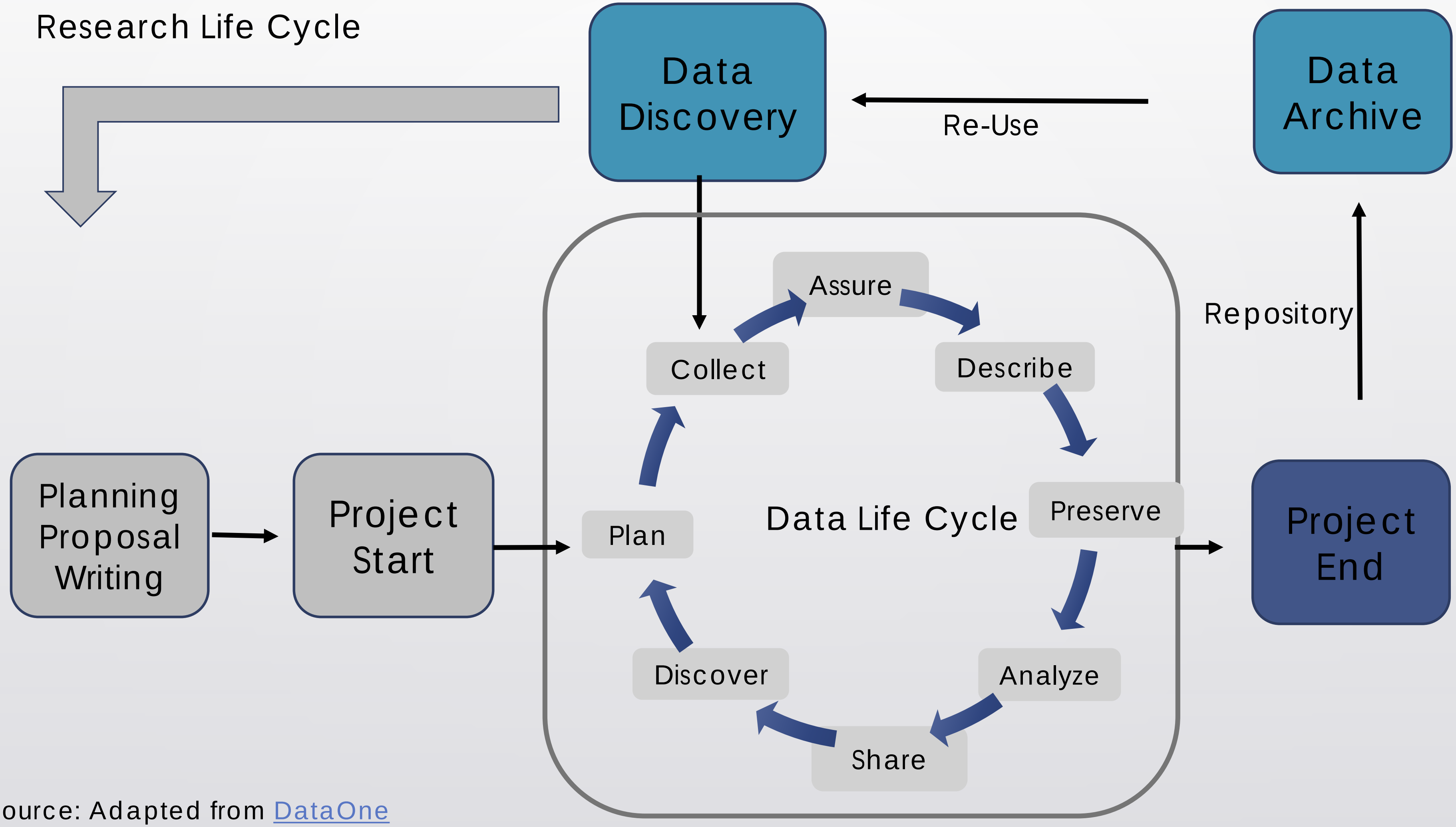
2 Describe fundamental processes of research data management (RDM)

3 Identify the components of a data management plan and construct a draft

why is data tlc important?

Source: NYU Health Sciences Library

research/data life cycles

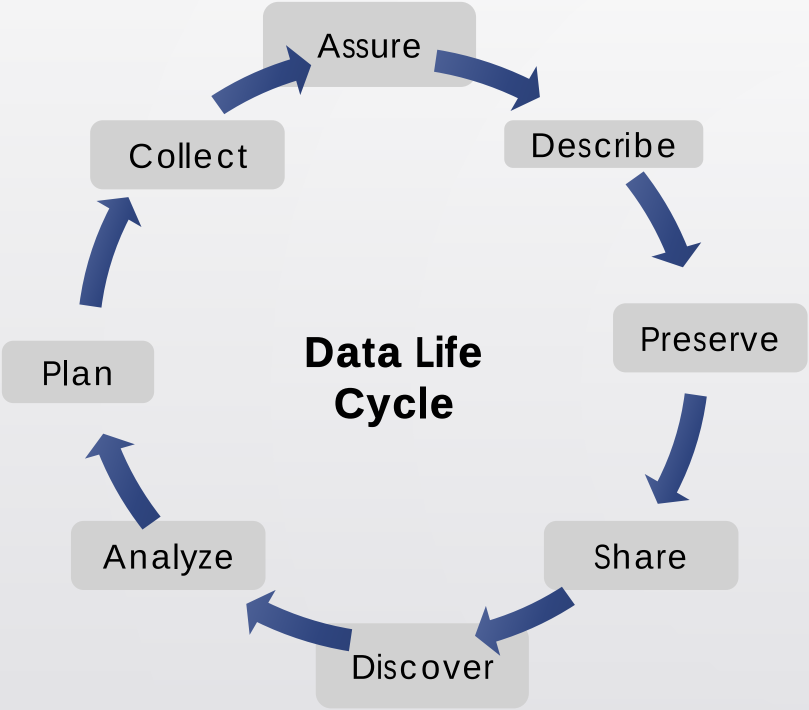


Source: Adapted from DataOne

how about your
data-related activities?

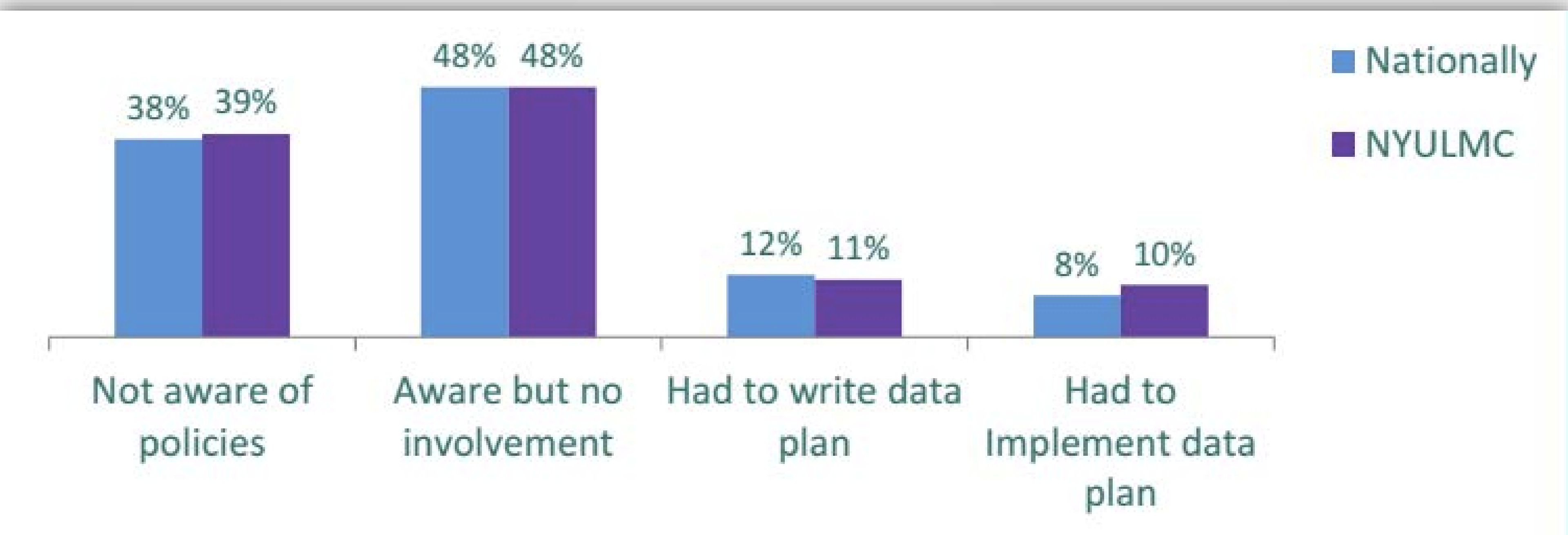
before the data

data management plan (dmp)



- ❑ Collection: ?
- ❑ Assurance: ?
- ❑ Description: ?
- ❑ Preservation: ?
- ❑ Sharing/Discovery: ?
- ❑ Analysis: ?

postdocs when surveyed asking “to what extent have you **dealt with** NIH data sharing regulations or NSF **data management plans**?”

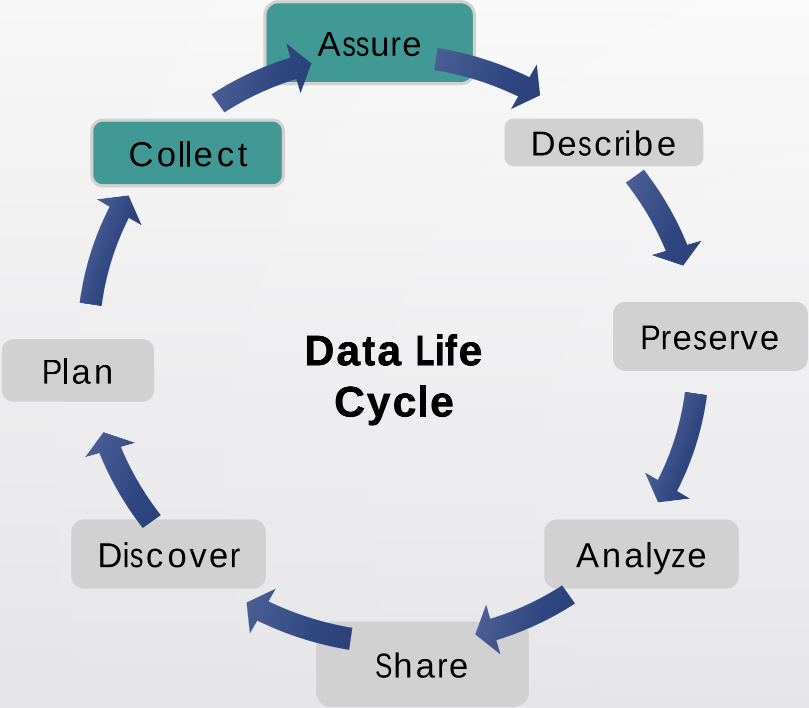


Source: NYU Medical Library

“Every minute you spend in planning saves 10 minutes in execution; this gives you a 1,000 percent return on energy!”

(Brian Tracy)

data collection



- ❑ Collection:
  - Where is the raw data?
  - Format and volume?
  - Methods?
  - Variables and units?

- ❑ Assurance:
  - Validation or standards (instruments/data)
  - Missing values?

Source: Digital Curation Centre

data standards in practice

Data standards are the rules by which data are described and recorded. In order to share, exchange, and understand data, we must standardize the format as well as the meaning.

Source: OHSU Open Educational Resources

different definitions for ‘data standards’

"Standards are documented agreements containing technical specifications or other precise criteria to be used consistently as rules, guidelines, or definitions of characteristics to ensure that materials, products, processes, and services are fit for their purpose."

(International Organization for Standardization (ISO))

Source: OHSU Open Educational Resources

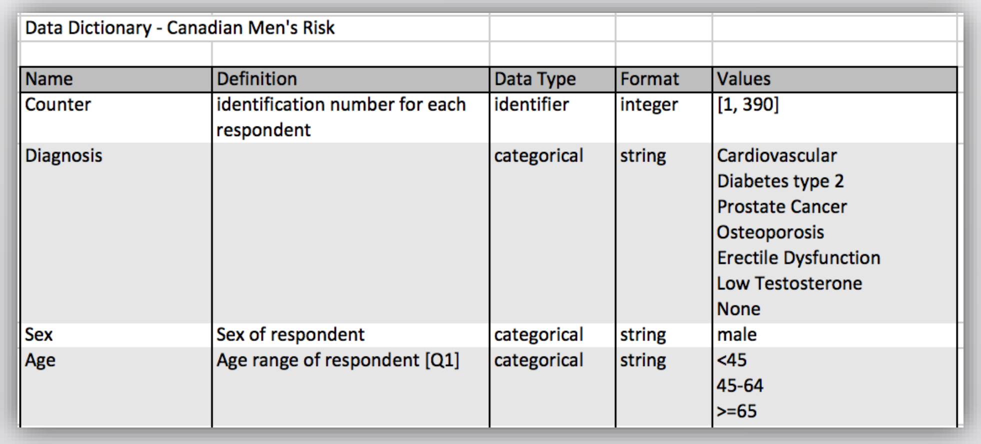
ways to standardize I

- ❑ Data content standards
  - ❑ Data content standards are a standardized, pre-defined terminology that is used consistently throughout a dataset.
  - ❑ For example, when describing a variable, use the same terminology throughout, such as male/female for gender (as opposed to M/F, men/women, etc.)

Source: OHSU Open Educational Resources

ways to standardize II

- Data dictionary
  - The purpose of a data dictionary is to ensure consistent terminology throughout a dataset. It contains a list of descriptive data types that are being collected.
  - Example:



Data Dictionary - Canadian Men's Risk

Name	Definition	Data Type	Format	Values
Counter	identification number for each respondent	identifier	integer	[1, 390]
Diagnosis		categorical	string	Cardiovascular Diabetes type 2 Prostate Cancer Osteoporosis Erectile Dysfunction Low Testosterone None
Sex	Sex of respondent	categorical	string	male
Age	Age range of respondent [Q1]	categorical	string	<45 45-64 >=65

Source: OHSU Open Educational Resources

ways to standardize III

- Data modeling
  - Data modeling is the a process that is typically performed before data collection. It involves identifying the user needs from the data set and determining the relationships among the data, and how the data should be best captured.

Source: OHSU Open Educational Resources

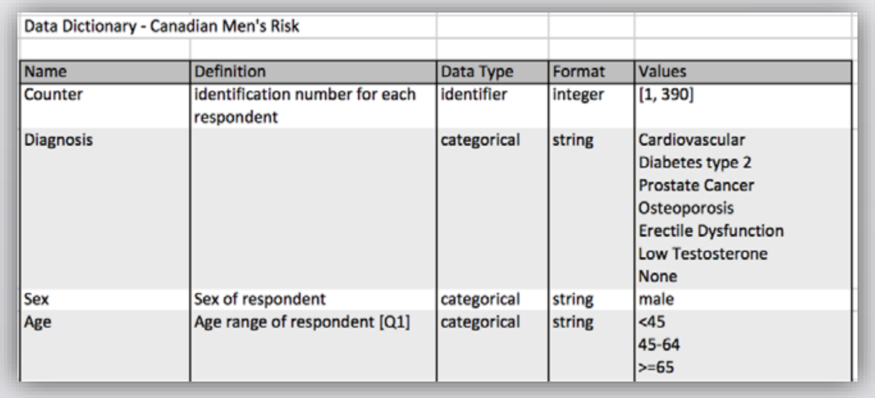
ways to standardize IV

- Ontology
  - A controlled vocabulary that defines the terms in a given domain or knowledgebase, and the relationships between those terms.
  - *Example*:
    - Gene Ontology
    - Mammalian Phenotype Ontology

Source: OHSU Open Educational Resources

ways to standardize V

- Metadata
  - Metadata is often referred to as “data about data”. Metadata is intended to describe other data, in order to clarify the meaning of a dataset. Metadata consists of both indexing terms (i.e., terms that can be used for search and retrieval of the data) and descriptive attributes.
  - *Example*:

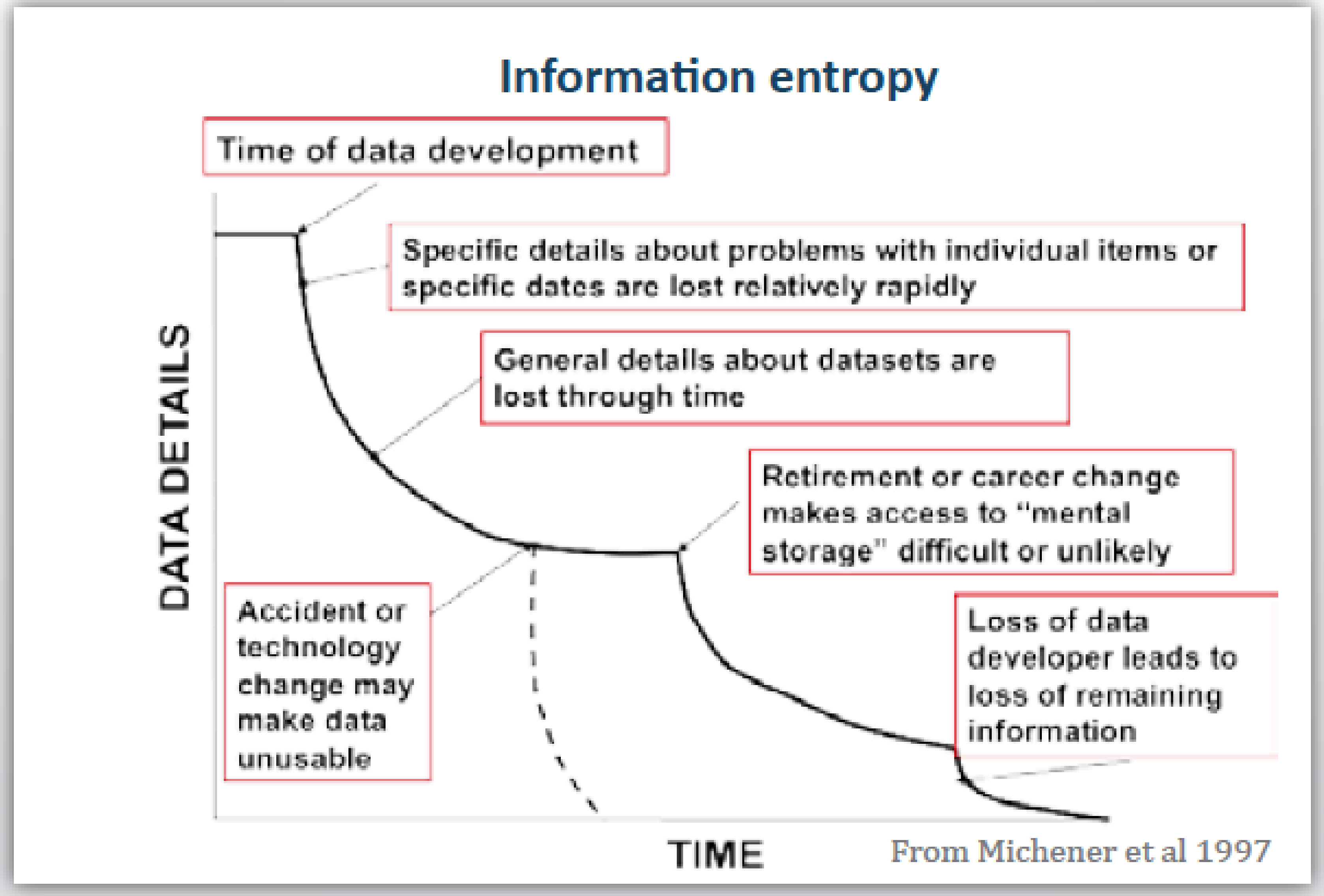


Data Dictionary - Canadian Men's Risk

Name	Definition	Data Type	Format	Values
Counter	identification number for each respondent	identifier	integer	[1, 390]
Diagnosis		categorical	string	Cardiovascular Diabetes type 2 Prostate Cancer Osteoporosis Erectile Dysfunction Low Testosterone None
Sex	Sex of respondent	categorical	string	male
Age	Age range of respondent [Q1]	categorical	string	<45 45-64 >=65

Source: OHSU Open Educational Resources

after the data



Information entropy

Time of data development

Specific details about problems with individual items or specific dates are lost relatively rapidly

General details about datasets are lost through time

Retirement or career change makes access to "mental storage" difficult or unlikely

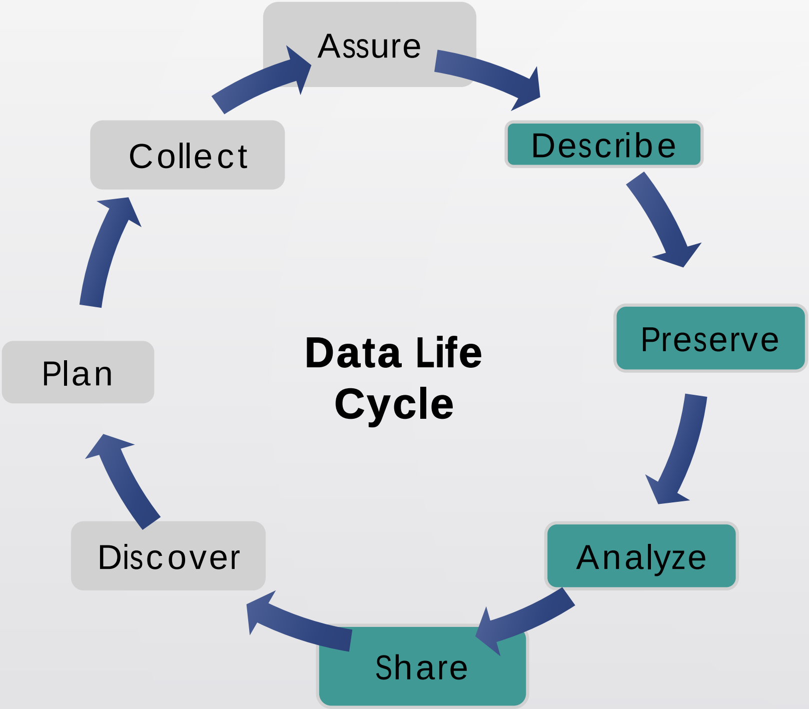
Accident or technology change may make data unusable

Loss of data developer leads to loss of remaining information

From Michener et al 1997

“Your closest collaborator is you six months ago, but you don’t reply to emails…”

data management plan (dmp)



- ❑ Description:
  - ❑ Descriptive details about the project & experiment

- ❑ Preservation:
  - Back up schedule?
  - 3 Back up locations?
  - How long to save it?

- ❑ Sharing:
  - Public repository?
  - Sensitive data?

- ❑ Analysis:
  - Tools (software)?

how do you describe your data?

Why are you doing this study?
Who is involved?
Where was the data collected?
What were the data collection processes
What do the values in the table mean?
What software is needed to view the data?
How should the data be cited?

README.txt
README.md

stable file formats

this	not that
.csv	.xls
.txt .html .rtf	.doc
.tiff .png	.jpg .gif

how will you store & backup the data?

- Computer/Laptop
- Network Drive
- External Devices
- Remote/Cloud
- Physical Storage (e.g. notebooks)

Source: OHSU Open Educational Resources

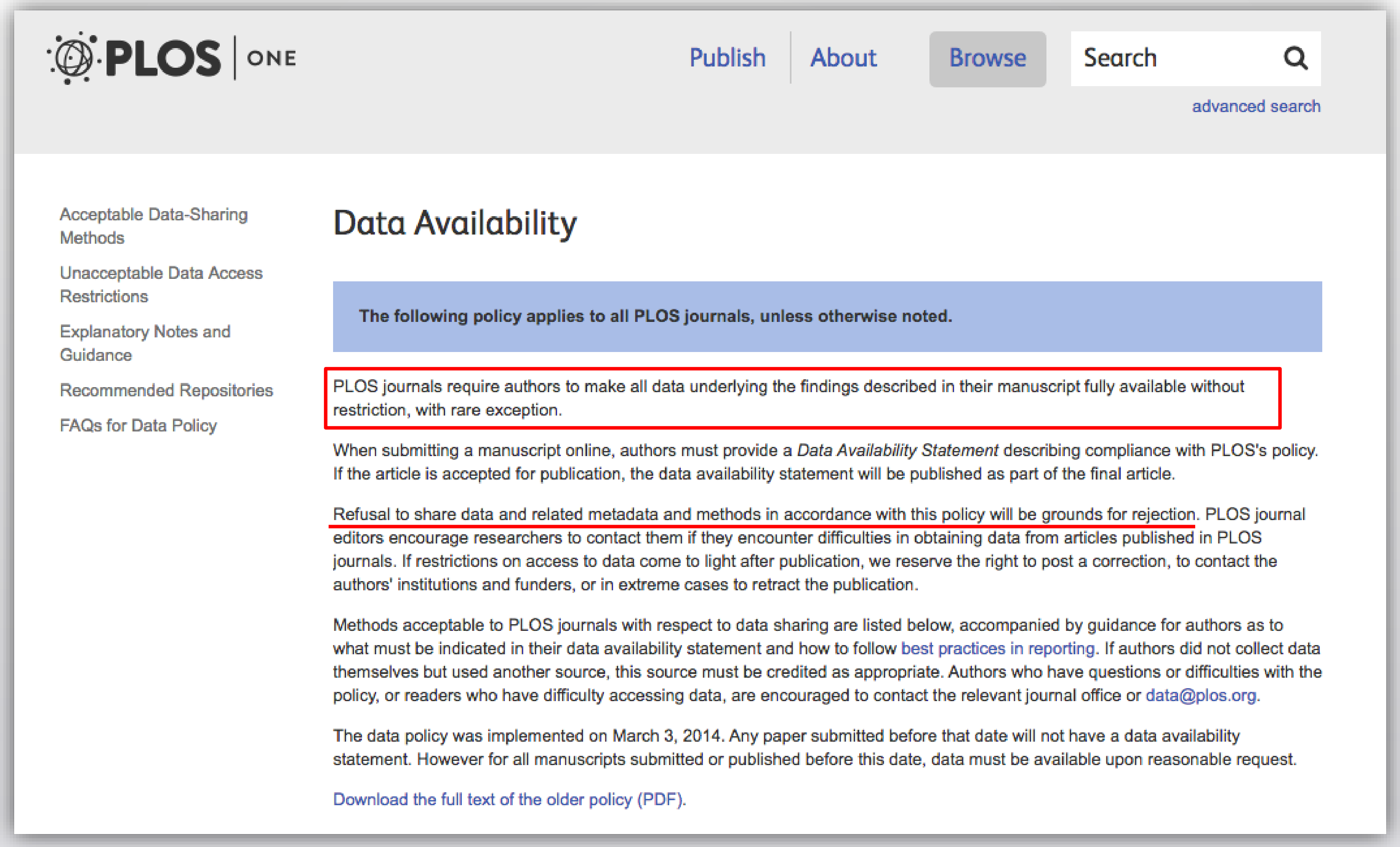
how will you store & backup the data?

Storage option	The good	The bad
Computer/Laptop	Convenient for active data	Easily lost/stolen Fail Manual backup
Network	Automatic backup and security	Access Capacity limitations
External devices	Low cost Portable Easy to use	Easily lost/stolen Fail
Remote/Cloud	Global access Collaboration	Security/privacy limitations
Physical storage (e.g. notebooks)	Convenient Tangible	Manual backup

Source: OHSU Open Educational Resources

how long will you keep the data?

preservation & sharing



PLOS ONE
Publish
About
Browse
Search
advanced search
Acceptable Data-Sharing Methods
Unacceptable Data Access Restrictions
Explanatory Notes and Guidance
Recommended Repositories
FAQs for Data Policy

Data Availability

The following policy applies to all PLOS journals, unless otherwise noted.

PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception.

When submitting a manuscript online, authors must provide a *Data Availability Statement* describing compliance with PLOS's policy. If the article is accepted for publication, the data availability statement will be published as part of the final article.

Refusal to share data and related metadata and methods in accordance with this policy will be grounds for rejection. PLOS journal editors encourage researchers to contact them if they encounter difficulties in obtaining data from articles published in PLOS journals. If restrictions on access to data come to light after publication, we reserve the right to post a correction, to contact the authors' institutions and funders, or in extreme cases to retract the publication.

Methods acceptable to PLOS journals with respect to data sharing are listed below, accompanied by guidance for authors as to what must be indicated in their data availability statement and how to follow best practices in reporting. If authors did not collect data themselves but used another source, this source must be credited as appropriate. Authors who have questions or difficulties with the policy, or readers who have difficulty accessing data, are encouraged to contact the relevant journal office or data@plos.org.

The data policy was implemented on March 3, 2014. Any paper submitted before that date will not have a data availability statement. However for all manuscripts submitted or published before this date, data must be available upon reasonable request.

Download the full text of the older policy (PDF).

Source: PLOS Author Guidelines

‘FAIR’ data principles

- □ What is ‘FAIR’ data?

  FINDABLE: persistent identifier, registered, rich metadata

  ACCESSIBLE: retrievable by ID, metadata accessible

  INTEROPERABLE: uses shared language, vocab is ’FAIR’, references other data

  RE-USABLE: accurate attributes, released with data use license, standards are met, provenance available

- □ Don’t forget about ‘FAIR’ metadata!

- □ Keep in mind for how you might craft a better plan for your managing your data…

Source: Force 11

data plans - why should we care?

incentives

- Maximize your research dollars by sharing and re-using data
- Maximize your research impact by sharing data
- Treat yo'self: general organization

Source: OHSU Open Educational Resources

funding agency mandates

❑ NIH

- ❑ Data sharing plan for new proposals
- ❑ Specific institutes and Centers have specific data sharing requirements

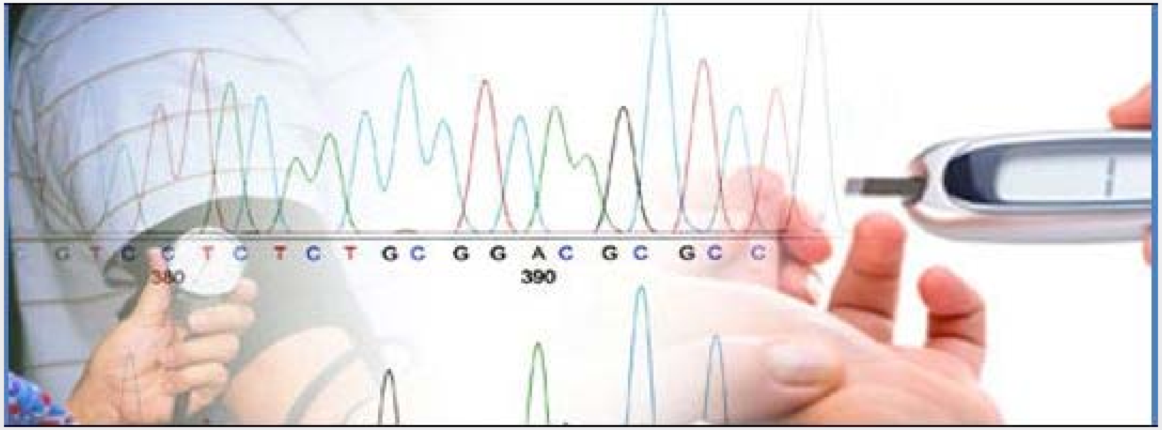
❑ NSF

- ❑ Requires a maximum 2 page data plan for data management and sharing of the products of research for all proposals

Source: [OHSU Open Educational Resources]()

funding agency mandates

- NIH/Genomic data



01101011 NCBI GenBank

dbSNP
Short Genetic Variations

MGI

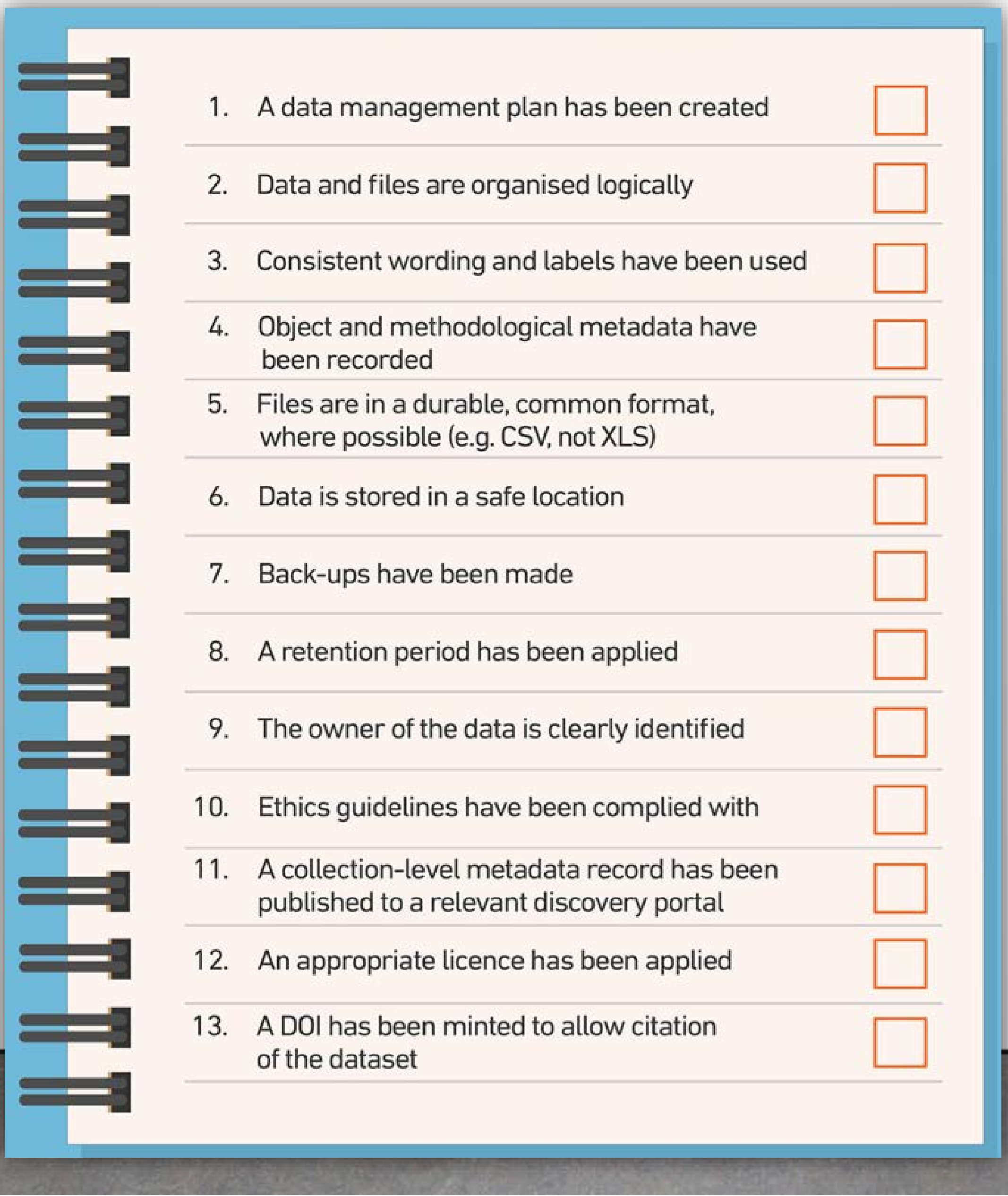
Source: OHSU Open Educational Resources

journal reporting requirements

- Reporting requirements vary
- Methods section is meant to contain enough information so someone could independently replicate your experiment
- Some journals are starting to increase their standards

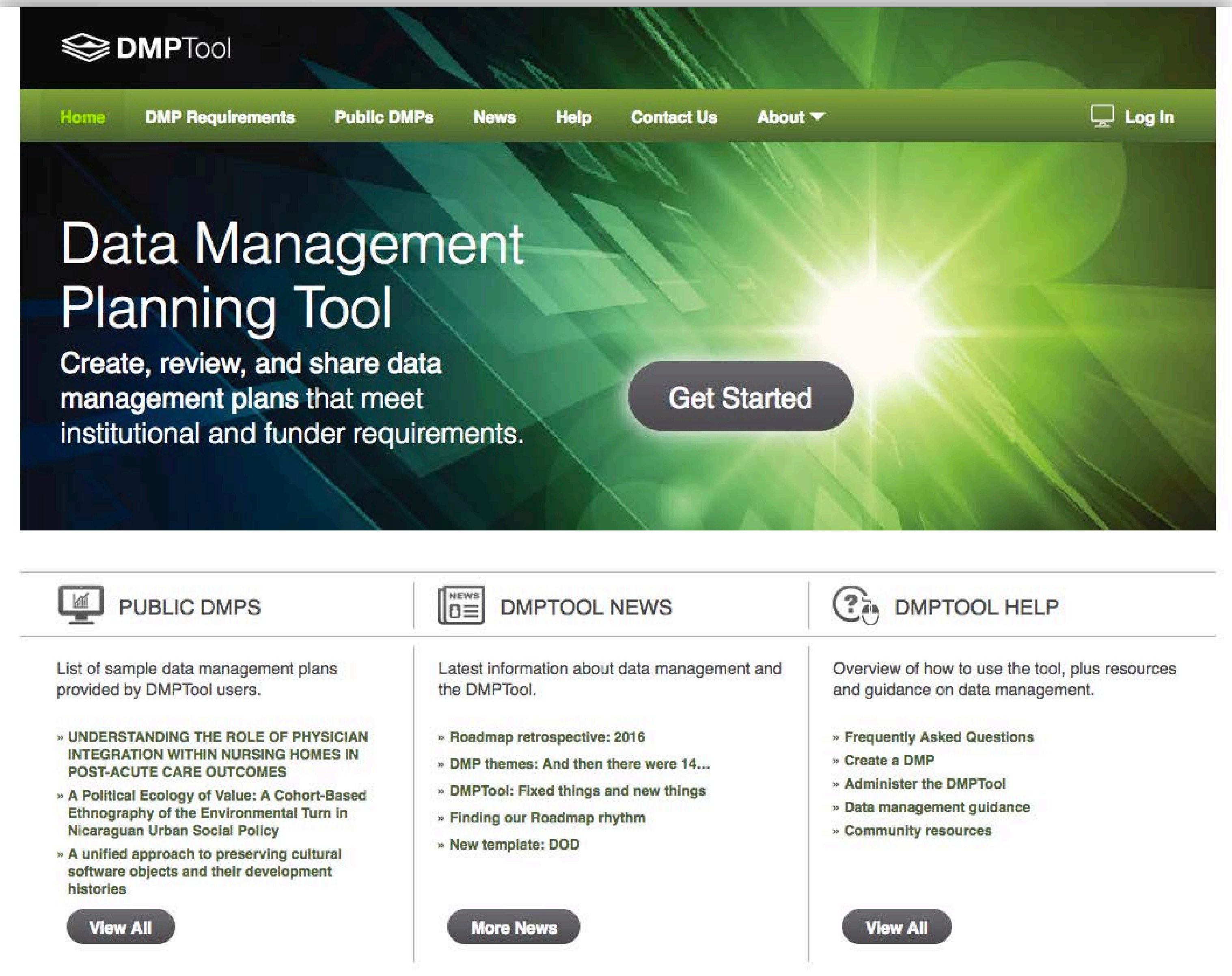
Source: OHSU Open Educational Resources

hang your checklist!



1. A data management plan has been created □
2. Data and files are organised logically □
3. Consistent wording and labels have been used □
4. Object and methodological metadata have been recorded □
5. Files are in a durable, common format, where possible (e.g. CSV, not XLS) □
6. Data is stored in a safe location □
7. Back-ups have been made □
8. A retention period has been applied □
9. The owner of the data is clearly identified □
10. Ethics guidelines have been complied with □
11. A collection-level metadata record has been published to a relevant discovery portal □
12. An appropriate licence has been applied □
13. A DOI has been minted to allow citation of the dataset □

dmp tool



Data Management Planning Tool

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

PUBLIC DMPS

List of sample data management plans provided by DMPTool users.

- » UNDERSTANDING THE ROLE OF PHYSICIAN INTEGRATION WITHIN NURSING HOMES IN POST-ACUTE CARE OUTCOMES
- » A Political Ecology of Value: A Cohort-Based Ethnography of the Environmental Turn in Nicaraguan Urban Social Policy
- » A unified approach to preserving cultural software objects and their development histories

DMPTOOL NEWS

Latest information about data management and the DMPTool.

- » Roadmap retrospective: 2016
- » DMP themes: And then there were 14...
- » DMPTool: Fixed things and new things
- » Finding our Roadmap rhythm
- » New template: DOD

DMPTOOL HELP

Overview of how to use the tool, plus resources and guidance on data management.

- » Frequently Asked Questions
- » Create a DMP
- » Administer the DMPTool
- » Data management guidance
- » Community resources

Source: DMPTool.org

your turn…

recap

1 Understand the elements of the data life cycle

2 Describe fundamental processes of research data management (RDM)

3 Identify the components of a data management plan and construct a draft

key resources

- ❑ OHSU Big Data to Knowledge (BD2k) Open Educational Resources (OERs)
- ❑ OHSU Library Data Webpage
- ❑ OHSU RDM Librarian (1:1 consulting)
- ❑ OHSU Library Workshops: Contact Us

questions?

Letisha R. Wyatt, PhD
wyattl@ohsu.edu
BICC 341
503.494.8627