RESEARCH DATA MANAGEMENT -- INTRODUCTION FOR ICORD

» Eugene Barsky and Paul Lesack

» UBC Library, May 2022

  » eugene.barsky@ubc.ca

  » paul.lesack@ubc.ca

» https://researchdata.library.ubc.ca

I ACKNOWLEDGE

I am grateful to the Musqueam people on whose unceded land I work, live and walk every day

AGENDA

» Define Research Data

» Research Data & The Data Life Cycle

» Data Management Plans

» Organizing Data: File Naming, README's & Metadata

» Storing & Protecting Research Data

» Data Repositories

WHAT IS DATA


Data
Information
Presentation
Knowledge
EpicGraphic.com

WHAT IS RESEARCH DATA

- Facts, measurements, observations, recordings, or records, produced during a research project.
  - Made of many file types, most disciplines have common types: Notes, spreasheets, images, designs, algorithms, diagrams, data files, etc.

RESEARCH DATA LIFECYCLE

» Start working with data with an end in-mind

WHAT IS RESEARCH DATA MANAGEMENT

» Processes applied throughout the lifecycle of a research project to guide the collection, documentation, storage, sharing, and preservation of research data.

» Portage Network RDM Primer, 2019

WHY PRIORITIZE RESEARCH DATA MANAGEMENT (RDM)?

» RDM helps organize your research data before, during, and after your research cycle.

» Makes responsibilities clear within research team & during team transitions

TRI-AGENCY FUNDING REQUIRES RDM

-As soon as Spring 2022, some Tri-Agency grants will require Research Data Management Plans in the application package.

Funding by Federal Funding Agency

Canada's Tri-Agencies (CIHR, NSERC, SSHRC) provide grants to support the direct costs of research. Grants are awarded through peer-reviewed competitions.

Tri-Agency	Total value	Grant Count	% of subtotal value
Canadian Institutes of Health Research (CIHR)	$149,155,370	1,208	55.25%
Natural Sciences & Engineering Research Council of Canada (NSERC)	$91,965,662	1,352	34.06%
Social Sciences & Humanities Research Council of Canada (SSHRC)	$24,924,548	624	9.23%
New Frontiers in Research Fund (NFRF)	$3,937,029	50	1.46%
Subtotal	**$269,982,609**	**3,234**	100%

GOOD RDM LOOKS LIKE:

» Clear distribution of team RDM responsibilities

» Accurate metadata about each research data file and variable

» A system for data storage, security, and backup

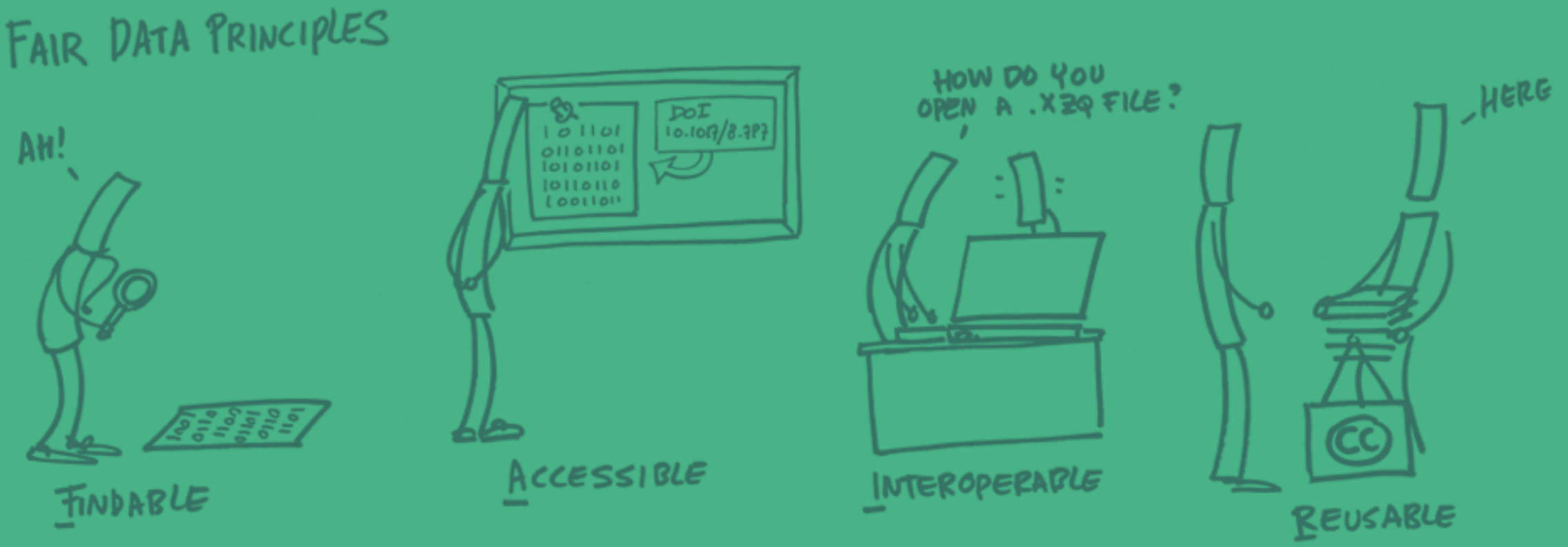
» Plan for data sharing, repository selection, and licensing for reuse

» Uniform and functional standards for naming and organizing data

GOOD RDM LOOKS LIKE:

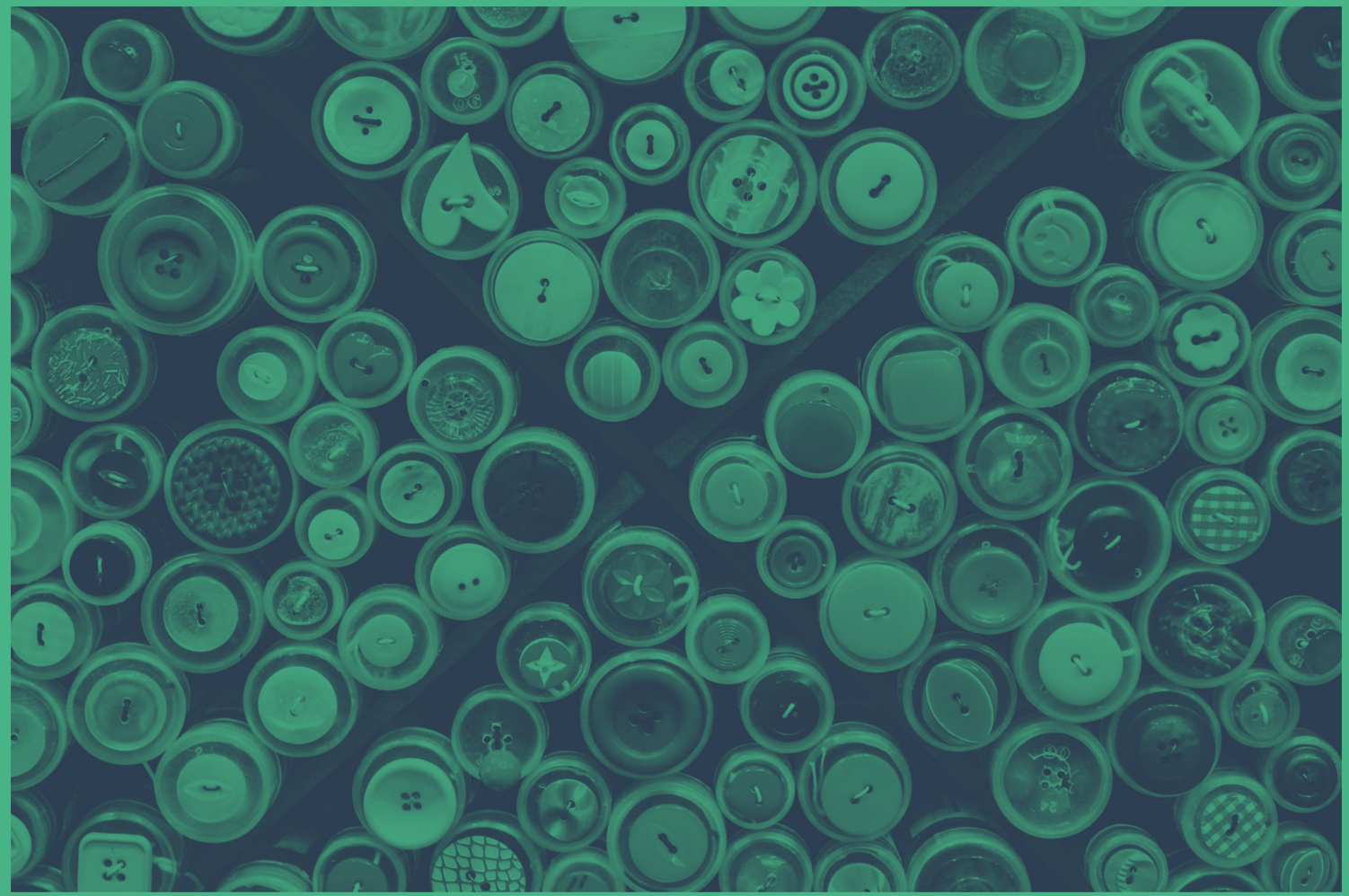
» My favourite example: Elizabeth Wolkovich's Lab Manual

FAIR DATA IS GOOD DATA



FROM HERE

TO HERE



PRACTICAL TIPS (#1) -- GET YOURSELF AN ORCID ID

» An ORCID iD is a free, persistent identifier (PID) that you can connect to your published research, including datasets.

» Like a DOI but for you across your entire career

» It makes it easy to gather your attributions, grants, positions, and more in one place

» Register for your ORCID iD here: https://orcid.org/

PLAN

THE DATA MANAGEMENT PLAN (DMP)

- Sets out how you will organize, store and share your research data at each stage in your project and after its conclusion.
- Incoming standard for funding agencies to require these in your grant applications.
  - Tri-Agency Research Data Management Policy

YOU ARE NOT ALONE

DATA MANAGEMENT PLAN (DMP) ASSISTANT BY PORTAGE

» Free and Bilingual - DMP Assistant

» Provides templates tailored to various disciplines

» UBC has custom templates for our researchers

» Templates guide the structure and creation of DMPs, focusing in on what's most important

DMP ASSISTANT

DMP EXEMPLARS

» Portage offers examples of completed DMPs.

» Exemplars - different disciplines

PRACTICAL TIPS (#2)

» Plan first

COLLECT

DATA COLLECTION
STOP PROBLEMS BEFORE THEY START

» Create a meaningful, organized system of variable names, file names, file structures

» Develop a directory structure (project, sample, etc.)

» Standardize your file naming system

DATA COLLECTION

» Document and share your process

» Write a comprehensive readme.txt outlining your variable names, applications, and notes that make the project replicable

DATA COLLECTION: FILE NAMING

» Aim for concise but descriptive file names, a consistent naming system, and a simple hierarchy.

» UBC Guide for File Naming

ELEMENTS THAT COULD APPEAR IN A FILE NAME

» Date of creation in ISO Standard YYYYMMDD format

» Name, initials or ID of the research group, institutional affiliation, editor, researcher

» A project name or code

» Short description of contents

» Location that produced the data

» Version number

» Format of the file, commonly seen as a file extension

FILE NAME EXAMPLES FROM SIMPLE TO COMPLEX

- A presentation that two people are working on:
  - ShortDescriptionEditorYearVersion
  - RDMWebinarNR2021V02
- Meeting notes for a group that meets regularly:
  - YYYYMMDDGroupDescription
  - 20200318RSpTMeetingNotes
- Tree points in standardized geographic areas in shapefile format
  - PointWesternRedCedarBCGS092G03220210719shp.zip

PRACTICAL TIPS (#3)

» Long, descriptive file names are better than short, confusing names, even if they take a long time to type

ANALYZE

PROCESSING YOUR DATA

» This can be iterative, in part or in entirety

» Keep an untouched, raw version of your data that is never changed.

» A DMP will ask you to consider where and how data is stored & backed up.

» Your code is also data

DATA STORAGE: THE CRUCIAL IMPORTANCE OF REDUNDANT SYSTEMS

- 3-2-1 backup rule:
  - Have at least 3 copies of your data
  - The here copy, your working copy
  - The near copy on a backup disk.
  - The far copy that can be accessed remotely.
  - Store the copies on 2 different media
  - Keep 1 backup copy offsite

DATA STORAGE

- How often shall I back up data?
  - As often as you edit your data, you should update your backups.
  - Personally I use FreeFileSync to sync in real time


entv
STUDIO STORIES
THE MOVIE VANISHES
NEW CLIP

UBC STORAGE RESOURCES

- UBC-IT
  - EduCloud Server Service - cost associated, need IT skills
  - Teamshare - Internal fileshare, costs per GB per year
  - OneDrive - 1TB of storage per user
  - Home Drive- personal storage only
  - SharePoint - powerful, but complicated to develop

UBC STORAGE RESOURCES

- UBC ARC
  - Chinook - Object storage application

UBC STORAGE RESOURCES

- Compute Canada
  - NetCloud - 100Gb free per user, similar to Dropbox
  - System storage - linux based for supporting high performance computing (HPC) analysis - up to 10Tb per research group.

UBC STORAGE RESOURCES

» UBC Online Storage Solutions Comparison Chart

STORAGE INTRICACIES

- Is it FIPPA compliant?
  - For Faculty & Staff: Campus Home Drives, TeamShare, Chinook, OneDrive, EduCloud
  - For Students: OneDrive
- Is any of your data is sensitive? Does it contain identifying information?

STORAGE INTRICACIES

» Is it free?

  » Yes: Home Drives up to 20GB, OneDrive, Chinook

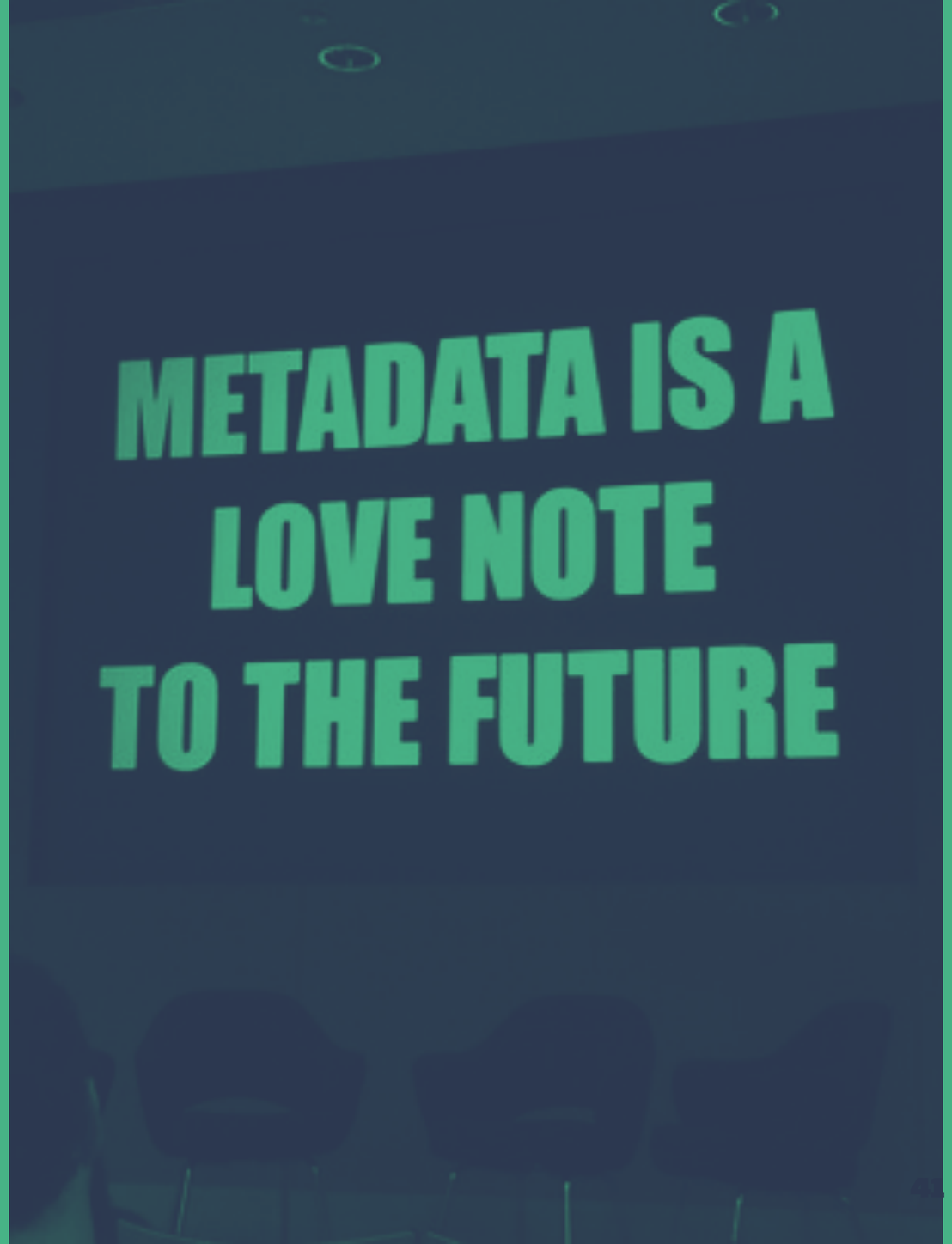
  » No: Team Share, EduCloud, SharePoint

PRACTICAL TIPS (#4)

» Backup your data. A lot.

WHAT IS METADATA? - DATA ABOUT DATA

- Metadata is descriptive data regarding a dataset, a study, a collection of files, and variables.
- Metadata answers questions like:
  - Who?
  - What?
  - When?
  - Where?
  - Why?
  - How?



METADATA ISN'T NECESSARILY FOR YOU

» As a researcher, you know what your data shows

» Metadata is for someone else, or for you in six months/20 years

TYPES OF METADATA

- Descriptive: content and context of your data at both the dataset level and data file level.
  - Title, Author, Methods, Sponsors

TYPES OF METADATA

- Administrative: information needed to use the data.
  - Software requirements, copyright
- Structural: how different data files relate to one another.
  - Information about the relationship between data files in a data set, file formats, variables

DATA DOCUMENTATION & README FILES

» README.txt files are the most basic tool for project documentation.

» Plain text (.txt) files that contain basic descriptive metadata about your project.

DATA DOCUMENTATION & README.TXT FILES

» README files accompany your data files throughout their life.

» UBC best practices for README files document

CONTENT OF README FILES

» Contact information for researchers

» Description of dataset & date of collection

» Use license for your data

» Methods & instruments used in the collection & processing of your data

» File structure and relations for the data set

» Explanations of codes, classifications, variables, and file names

This readme.txt file was generated on YYYY-MM-DD by INSERT RESEARCHER NAME

GENERAL INFORMATION

1. Title of Dataset:

2. Author Information

Name:
Institution:
ORCID:
Email:

3. Description of Dataset

4. Date of data collection:

5. Geographic location of data collection:

SHARING/ACCESS INFORMATION
1. Licenses/restrictions placed on the data:

DATA & FILE OVERVIEW
1. File List:

VARIABLE DESCRIPTIONS

Variable:	The variable indicates:
Included	whether the participant's data was included in the analysis (1 = yes; 0 = no)
Experimenter	which of the four male experimenters conducted the session
T1_Time	time of day (24h) at commencement of saliva sample 1 collection
GameStart	elapsed time (hh:mm:ss) between the start of saliva sample 1 collection and the first spin
EndEarly	whether the participant voluntarily discontinued gambling (1 = yes; 0 = no)

NOT EVERYBODY KNOWS...

» Define all jargon, “common knowledge,” and abbreviations

» Your data and README.txt files may be utilized by researchers outside of your field of study who may not have the same understandings

» Knowledge changes over time, don’t take understandings for granted

```
AGE - Participant’s age
SAI1 -
SAI2 -
SAI3 -
PTW - Minutes spent on public transport/week
```

PRACTICAL TIPS (#5)

» Write your documentation so that you have a self-contained data package for a non-specialist.

SHARE AND PRESERVE

SHARE & PRESERVE YOUR DATA FOR REUSE

» Much easier when you've maintained your RDM and implemented your DMP

» Effectively boxing up some of your research data

» All documentation and metadata should be complete at this stage

SHARE & PRESERVE YOUR DATA FOR REUSE

» May require converting to stable file-formats instead of proprietary formats

  » Tabular data as .CSV/.TSV, text files as .TXT or .PDF, etc.

LONG-TERM STORAGE CONSIDERATIONS

- Where do you plan to hold the data long term?
- What forms of the data are you sharing?
  - Raw, processed, analyzed, and final are all options
  - Do the funding agency or publisher mandate any of these for you?

LONG-TERM STORAGE CONSIDERATIONS

» What license do you plan to apply?

» Personally I like the CC-0 and CC-BY for research data

PRACTICAL TIPS (#6)

» Data licenses aren't optional. Pick one which meets the needs of you and your intended audience

DATA REPOSITORIES

- Benefits of a repository:
  - Persistent, unique identifier, e.g. DOI
  - Backup and preservation
  - Data citation
  - License
  - Version control
  - Data integrity checks

UBC REPOSITORIES

» UBC Dataverse Collection - An open source application to publish, share, reference, cite, extract and analyze research data. (Preferred)

» FRDR - The Federated Research Data Repository, is a Canadian national research data repository.

» Dryad - Dryad is an international, multi-disciplinary data repository that supports access to data underlying published literature. UBC is a Dryad institutional partner.

UBC REPOSITORIES

» Where Should I Deposit My Data – Quick Guide

UBC DATAVERSE COLLECTION

» UBC CWL integration

» Collaboration among Canadian institutions

  » Data is stored in Canada

» Limits individual files to 3 GB each but accepts all data formats

» Mints DOIs for datasets

» Discovery by Google, Summon, FRDR, DataCite

DRYAD

» UBC CWL integration.

» Data curation review

» Digital Object Identifiers (DOIs) at the dataset level.

» No fees for basic data deposits - UBC Library is an institutional member

» If your files are <3GB, your data set goes into Dataverse anyway

» Guides and instructions

FRDR

» Data curation review

» Digital Object Identifiers (DOIs) at the dataset level.

» No fees for basic data deposits - UBC Library is an institutional member

» Guides and instructions

» Support for large files

THE CIRCLE OF REPOSITORIES

» FRDR harvests from Dataverse

» Dataverse harvests from Dryad

HOW TO CHOOSE

For most purposes, Dataverse is a good choice

» Version control and persistent identifiers

» Obvious UBC branding

» Can (optionally) have collections by

  » organization (ie, ICORD)

  » PI/Lab

CHALLENGES TO DEPOSITING, SHARING REUSING DATA

- Lack of metadata
  - Include sufficient metadata (see tip #5)
- Proprietary, obsolete file formats
  - Use preservation-friendly, open file formats
- Copyright, intellectual property rights unclear - No license
  - Encourage open access to data
- Privacy, ethical concerns
  - Obtain consent for data sharing and secondary use of data
  - Check with Research Ethics Boards

RDM PEOPLE AT UBC

» Research Data Services Librarian, Eugene Barsky

» Paul Lesack, Data/GIS analyst

  » research.data@ubc.ca

» UBC Library's Research Data Management Site

» Advanced Research Computing (ARC) at UBC

» Research Commons Consultations & Workshops

RDM GUIDES AT UBC

» UBC Library Data Guide

» Good Enough Research Data Management - A Very Brief Guide

» File Naming Guidelines

» Creating A README For Your Dataset

UPCOMING RDM WORKSHOPS BY UBC LIBRARY

» UBC Library Research Commons hosts workshops on topics like citation management tools, systematic reviews, metadata, GitHub, R, NVivo, and more.

» Upcoming Research Commons Workshops

THANK YOU FOR JOINING US!

- research.data@ubc.ca
- eugene.barsky@ubc.ca
- paul.lesack@ubc.ca

IMAGE CREDITS

» Slide 15 - OpenAIRE

» Slide 16, 17 - Unsplash

» Slide 28 - XKCD

» Slide 37 - Flickr

» Slide 47 - Flickr

» Slide 50 - Flickr