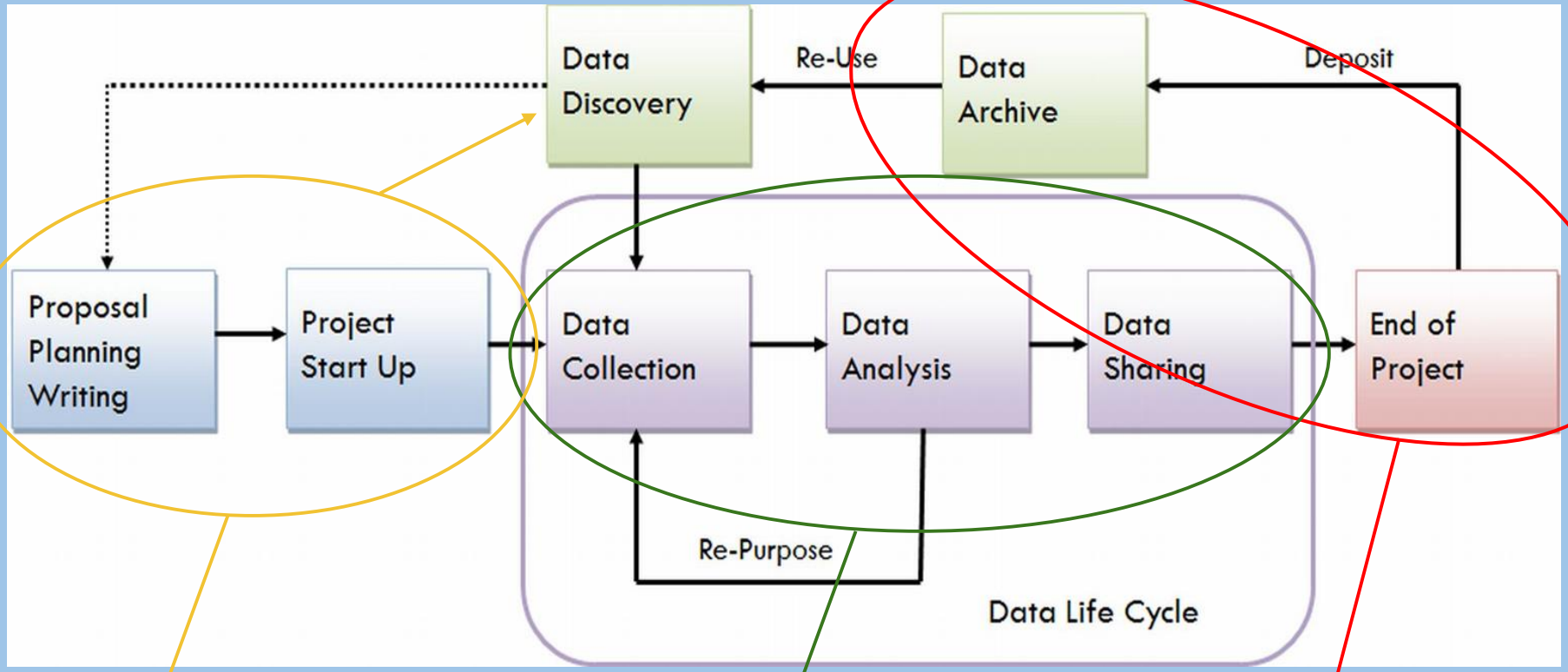
(Some) Research Data Management Best Practices

James Doiron
RDM Services Coordinator, UofA Libraries
Academic Director, UofA Research Data Centre
Co-Chair, Portage DMP Expert Group

Shifting Horizons II: Realities of RDM Services
University of Ottawa
February 20, 2020

(Some) RDM ‘Best Practices’: A bird’s eye view

1. Overarching principles: FAIR; CARE; OCAP



2. Pre-Research:
   - Including RDM into funding applications
   - Funder requirements
   - Data Management Plans
   - Data discovery
   - Participant consent & Information Letters
   - Ethics applications

3. During Research:
   - Primary data collection
   - Data storage
   - Data transferring
   - Data access
   - Data documentation/metadata
   - File naming conventions

4. Post-Research:
   - Publication requirements
   - Data deposit

1. Overarching Principles

FAIR

FAIR is a set of guiding principles focused towards making data **F**indable, **A**ccessible, **I**nteroperable and **R**eusable:

- **Findable** - Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.
- **Accessible -** Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.
- **Interoperable** - Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- **Reusable -** Data and collections have clear usage licenses and provide accurate information on provenance.

*Key Reading: Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*3:160018 doi: 10.1038/sdata.2016.18 (2016).

FAIR Principles: Key Resource

go-fair.org (https://www.go-fair.org/fair-principles/)

- Detailed information across the FAIR principles
- Implementation Networks
- News
- Event
- Resources!

1. Overarching Principles

CARE

CARE is a set of guiding principles for Indigenous data governance:

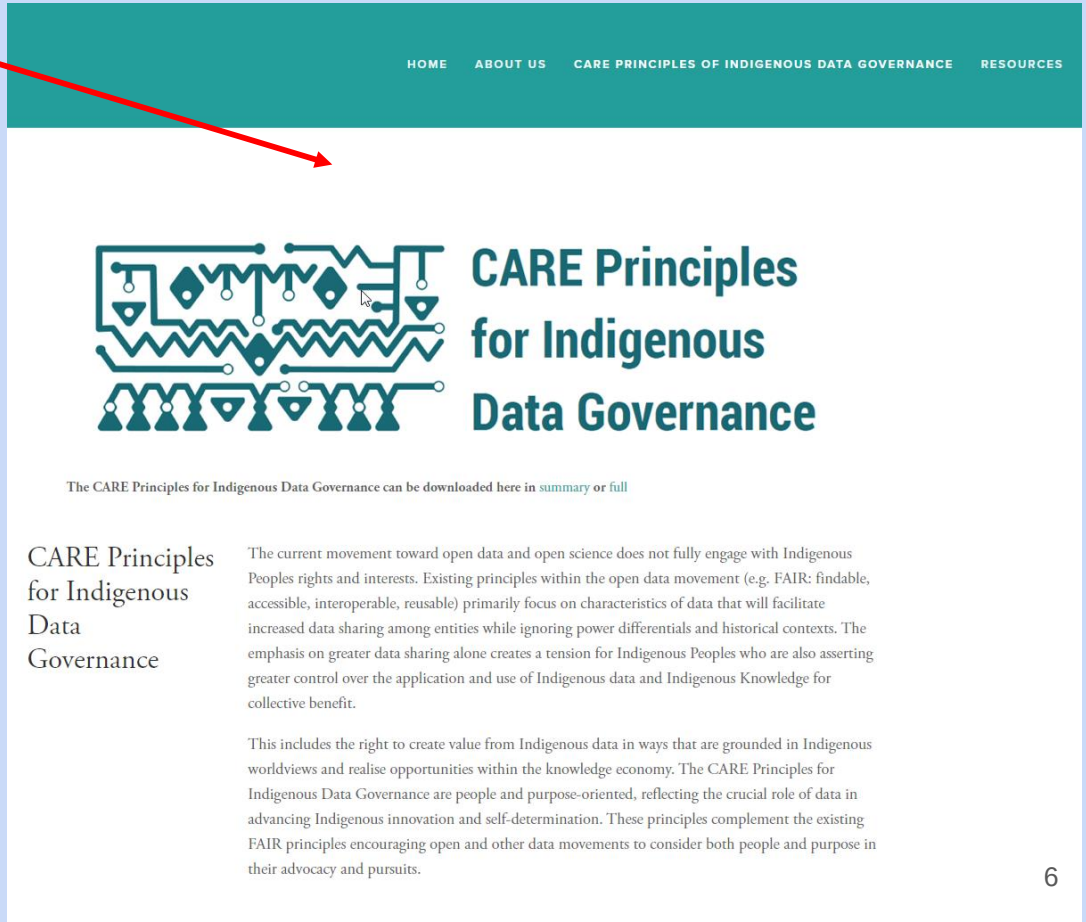
- **C**ollective benefit for inclusive development and innovation, improved governance and citizen engagement, and equitable outcomes
- **A**uthority to control - Recognizing rights and interests, data for governance, and governance of data
- **R**esponsibility for positive relationships, expanding capability and capacity, and Indigenous languages and worldviews
- **E**thics for minimizing harm and maximizing benefit, justice, and future use of data

*Key Readings available at: https://www.gida-global.org/resources

CARE Principles: Key Resource

gida-global.org/care

- Detailed information across the CARE principles
- Foundational readings & publications
- News
- Events
- Resources!



1. Overarching Principles

OCAP

The First Nation Principles of OCAP are a set of standards that establish how First Nations data should be collected, protected, used or shared:

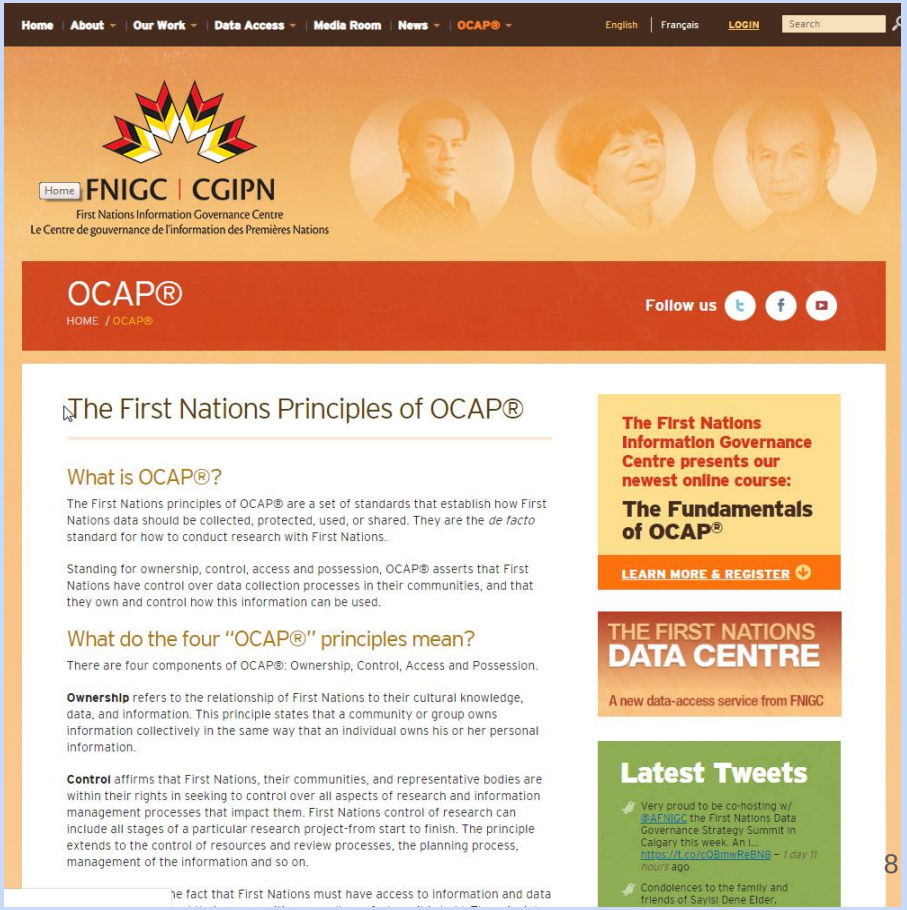
- **Ownership:** refers to the relationship of First Nations to their cultural knowledge, data & information - a **community/group collectively owns information** in the same way that an individual owns his/her personal information
- **Control:** affirms that **First Nations communities have rights in seeking control over all aspects of research** - from start to finish - that impact them. This extends to control of resources and review processes and management of information.
- **Access: First Nations must have access to information and data about themselves and their communities** regardless of where it is held, and have the **right to manage and make decisions** regarding access to their collective information.
- **Possession:** Refers to the physical control of data - the **mechanism by which ownership can be asserted and protected.**

"OCAP® is a registered trademark of the First Nations Information Governance Centre (FNIGC)"
www.FNIGC.ca/OCAP

OCAP Principles: Key Resource

First Nations Information Governance Centre
(https://www.FNIGC.ca/)

- Fundamentals of OCAP online training program
- FNIGC data online
- First Nations Data Centre (data by request)
- First Nations surveys (i.e., regional health, early childhood, education, labour, oral health)
- FNIGC online library



2. Pre-Research

Funder requirements

Awareness of funder requirements helps to identify:

- specific supports needed;
- collaborative opportunities;
- RDM supports to leverage

When involved in funded research projects, know <u>who</u> the funder is and:

1) What general RDM related policies they may have;

2) If there are any RDM related requirements pertaining to the specific call for funding

2. Pre-Research

General RDM Funder requirements

Example:
'*Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans*' (TCPS-2)

2. Pre-Research

General RDM Funder requirements

Example: Tri-Agency Open Access Policy on Publications

Policy Summary

Researchers awarded funding from January 1, 2008 onwards from CIHR are required to adhere with the following responsibilities:

- ensure that all research papers generated from CIHR funded projects are freely accessible through the Publisher's website or an online repository within 12 months of publication;
- deposit bioinformatics, atomic, and molecular coordinate data into the appropriate public database (e.g. gene sequences deposited in GenBank) immediately upon publication of research results;
- retain original data sets for a minimum of five years (or longer if other policies apply);
- and acknowledge CIHR support by quoting the funding reference number in journal publications.

Tri-Agency Open Access Policy on Publications

Adhering with the policy – Open access publications

For journal publications, there are two ways to adhere with the policy:

- Submit your manuscript to a journal that offers immediate open access or offers open access to the paper on its website within 12 months of publication.
- Submit your manuscript to a journal that does not offer open access, but will permit you to archive the peer-reviewed manuscript in a central or institutional repository within 12 months of publication.

The SHERPA/RoMEO database contains a searchable listing of journal publisher's copyright and self-archiving policies which will help researchers to determine journals that adhere with CIHR policy.

2. Pre-Research

Specific Funder requirements

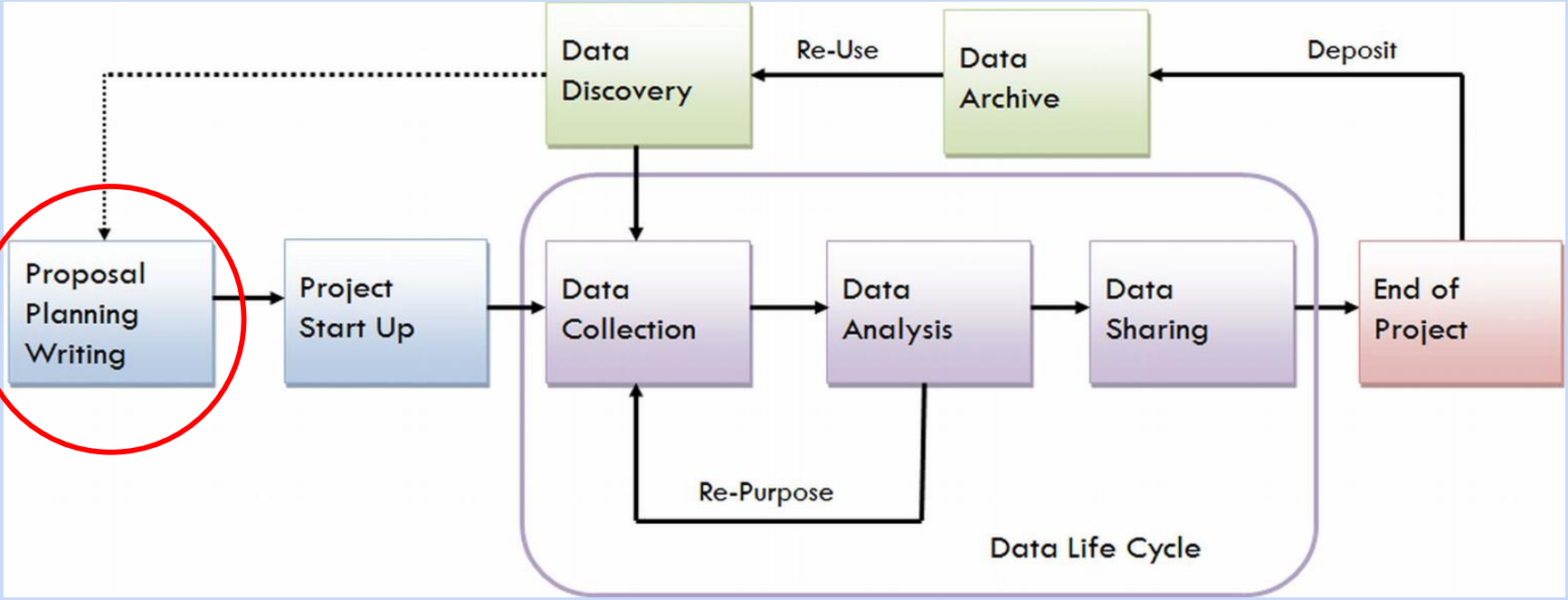
Example: CRAft Digital Research Archive grants

- Research data management plan with a focus on data accessibility and stewardship. ***Please note that the RDM plan is not considered part of the 5-page proposal and should be included as an attachment. We recommend using the Portage DMP Assistant to generate an RDM plan.***

Including RDM into funding applications

Including RDM within funding applications can help to strengthen funding applications by identifying:

- areas where essential RDM support may be needed
- collaborative opportunities *before* the research begins
- specific RDM supports that research projects may leverage



2. Pre-Research

Including RDM into funding applications

Example RDM statement:

Research Data Management: The University of Alberta Libraries system will provide research data management training and support for project researchers on a one-on-one and group basis (including HQP), host project research data in Dataverse, UAlberta's data repository, and host project papers and publications, learning objects, digital images, etc. in its open-access Education & Research Archive.[i] Data management will extend beyond the project itself to ensure sustainability of the data for future researchers. We will form a *Research Data Management Committee* (reporting to the *EC*; see *Governance*) and use the *Portage DMP Assistant*[ii], a web-based open source application, to develop a *Research Partnership Data Sharing Agreement* at the beginning of the formal partnership. One of the major outcomes of this project will be well-documented, well-preserved **data sets** which can be used by future researchers and are themselves a form of scholarship.

**Special thanks to Dr. Carla Peck from the Faculty of Education, UofA, for permission to use this text from her recently successful application for SSHRC Partnership funding - "Thinking Historically for Canada's Future"*

2. Pre-Research

Data Management Plans

We will be covering DMPs in-depth this afternoon!

2. Pre-Research

Data Discovery→ Existing data?

When developing a research project it can be useful to explore whether there are existing data that can help to answer research questions or to strengthen funding applications.

Existing data can:

- Reduce need for primary data collection (*data repurposing*)
- Guide development of research questions
- Identify ‘data gaps’ (*need for primary data collection*)
- Inform the development of data collection instruments

2. Pre-Research

Data Discovery→ Existing data?

Some data source examples:

- UOttawa Libraries
- Statistics Canada Data Liberation Initiative (DLI)
- Canadian Research Data Centre Network (CRDCN)
- Canadian Institute for Health Information (CIHI)
  (see: Graduate Student Data Access Program → no cost data)
- First Nations Information Governance Centre (FNIGC)

2. Pre-Research

Participant consent & Information letters

Research involving human participants require informed consent

Information letters must describe *how data are handled during active phases and beyond*, are outlined.

What can/can't be done will in part be determined by what is said in:

- Participant information letters/consent forms
- Approved ethics applications

NOTE: It can be very difficult - or even impossible - to go back to participants in efforts to revise their consent, so getting it right from the get-go is important!

2. Pre-Research

Participant consent & Information letters

Outline such things as:

- project background
- purpose of the study
- study procedures
- benefits/risks
- data preservation/destruction
- security/confidentiality
- voluntary participation
- freedom to withdraw

***FUTURE USE OF DATA:**

Participant consent is required in order for data to be used beyond the scope of the immediate project

Example statement:

*"By participating in this research I hereby give consent for my *de-identified* information to be used for research purposes beyond this immediate project"*

2. Pre-Research

Ethics applications

An ethics application addresses such things as:

Research design/methodology, risks/benefits, security/confidentiality, participant information, informed consent, data sensitivity, data collection, & data storage, retention and disposal

Most research project involving human participants require ethics approval

Multi-institutional/regional projects require multiple ethics approvals

The UOttawa Office of Research Ethics & Integrity offers essential supports & services to help guide and support the ethical management of your research data

2. Pre-Research

Ethics applications

***FUTURE USE OF DATA:**

If there is potential for future use of data outside of the immediate project then this should be clearly stated within the ethics application

Example statement:

*"There are no plans to destroy these data. Data will be securely stored *enter details of storage methods'*for *i.e., the minimum 5 years**

**De-identified* data may be deposited into an institutional repository (e.g., UOttawa's Dataverse) for discovery and possible repurposing. Any future use of these data outside of the immediate research project will occur only with all ethical and contractual obligations met"*

3. During Research

Data collection: immediate storage

A safe definition of ‘storage’ = 48+ hours

Storing data on mobile devices is not considered best practice

Any electronic devices used for collecting/storing data should always be encrypted *(i.e., laptops, digital voice recorders, tablets, etc)*

Best Practice:

- Clear & succinct data collection policies and protocols that *define when and how data are transferred off of data collection devices*
- Be aware of any data storage policies imposed by institutions, funders, data providers, etc.

3.

$1M study at U of M paused after personal health info of 420 participants is breached

Study was led by Peter Jones, who was suspended by the university last year

Kristin Annable · CBC News · Posted: Mar 06, 2019 5:23 PM CT | Last Updated: March 6

Any electronic devices used for collecting/storing data should always be encrypted *(i.e., laptops, digital voice recorders, tablets, etc)*

It informed them their health information was "not handled, stored or secured properly." According to the letter from the university's access and privacy office, an audit of the program found data was being stored off-site, without encryption, allowing the potential for it to be accessed by a third-party company.

Best Pra

- Clea *when and how data are transferred on or data collection devices*
- Be aw ders, data providers, etc.

The announcement of the PHIA breach means all the tests and data from the 420 participants will be destroyed and will no longer be used in the study.

3. During Research

Data collection: ‘longer term’ immediate storage

Sometimes it is necessary to store data on electronic devices - i.e., laptops, portable hard drives - for longer periods (collecting data in rural/remote areas)

Beyond security risks, these also introduce risk of data loss and/or corruption

Best Practice:

If the use of laptops/desktops/hard drives is deemed necessary for longer term data storage use the 3-2-1 rule:

At least 3 independent copies of your data
Store copies on 2 different types of media
Keep 1 backup copy offsite

3. During Research

Data transferring

Transferring of data is a critical stage of the data collection process

Regardless of whether data are collected from primary or secondary sources, transferring of data is a necessity

Some Risks:

Data transfers may occur:

- from field (real world settings)
- from data providers
- between researchers
- between researchers & stakeholders

3. During Research

Data transferring

Transferring of data is a critical stage of the data collection process

Regardless of whether data are collected from primary or secondary sources, transferring of data is a necessity

Some Risks:

3. During Research

Data transferring: Risks

Definite don’ts: e-mail, drop-box, unencrypted devices

Typical do’s: Secure FTP; MS Sharepoint; secure extranets

Best practice:

- Identify data transfer methods that you will use before the research begins
- Talk to your local IT support to identify secure methods available

Data storage - Cloud Services

What is ‘Cloud Storage’

- Physical storage typically spans multiple servers (sometimes in different locations)
- Data are easily available and remotely accessible→ typically 24/7
- Best practice → built in physical, technical, & administrative safeguards

3. During Research - Cloud Services

Do you know where your cloud is?

3. During Research - Cloud Services

Do you know where your cloud is?

‘Clouds’ are clusters of servers → servers need to live somewhere

Find out:

- Where servers are physically located before using a cloud service → local, provincial, Canada?

- What security policies and procedures are in place→ disaster recovery, back-ups, etc

→ **Research Portal**

- **Rapid Access Service** → cloud space for researchers; up to 1TB at no cost
- **Resource allocation competitions** → for greater needs
- **High performance computing**
- **Big data transfers**→ i.e., TBs of data
- **Portals for specialized software & tools**
- **Data storage & back-up**

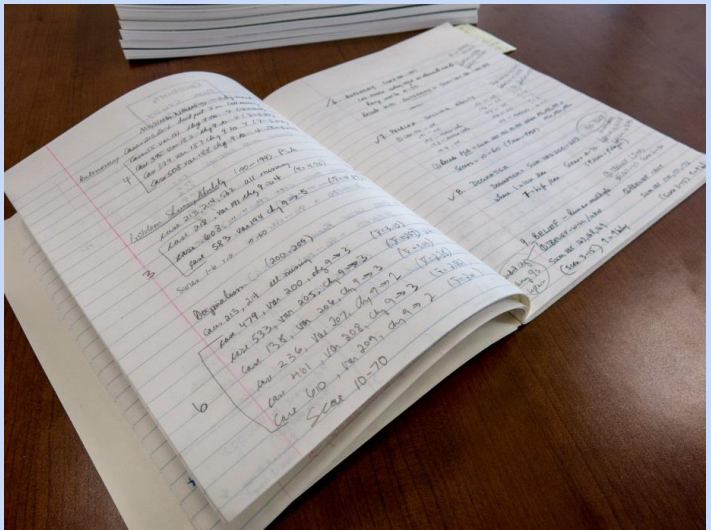
3. During Research→ **Data documentation & metadata**

‘Metadata’ essentially refers to ‘data about your data’

Descriptive information that describe your data, as well as to help others (and machines) to locate it and make it readable and useable

Some key examples of metadata documentation can include:

- Data codebooks/dictionaries
- Data management and processing protocols
- Readme files
- Analytic plans
- Code



3. During Research→ **Data documentation & metadata**

There are three main types of metadata:

- **Descriptive metadata** describes your research study & data in order to help facilitate their identification and discovery. *(i.e., author, study title, how data were created, keywords, etc)*

- **Structural metadata** consists of information relating to the type(s) of data, versions, and how the data relates to other associated digital materials

- **Administrative metadata** includes information such as file type and size, when and how the data were created, and other technical properties

3. During Research→ **Metadata standards**

Metadata standards help to ensure that different systems, computers and software are able to exchange information→ facilitates the discovery, access, and use of your data by others.

There are **many** different types of metadata standards such as:

- **Data Documentation Initiative (DDI)** (Social Sciences; Arts & Humanities)
- **Dublin Core** (General)
- **Investigation/Study/Assay tab-delimited (ISA-TAB)** (Engineering; Life Sciences)
- **ISO 19115** (Physical Sciences & Mathematics);
- **Statistical Data and Metadata Exchange (SDMX)** (Social & Behavioural Sciences)

A comprehensive overview of metadata standards, including which are most applicable for use within various research disciplines, is available at the Research Data Alliance Metadata Directory

3. During Research - File naming

Having clear and standardized file naming helps to support:

- **Organization**
- **Quality assurance**
- **File versioning**
- **Collaborative use**
- **Data analysis**
- **Dissemination**
- **Preservation & archiving**
- **Staff/Student training**

Elements of file names can include:

- Project name/acronym
- File version
- Data type
- Participant codes/pseudonyms
- Geographic location
- Context information
- Date information
- Interviewer codes/initials

Working with data: File versioning

File versioning is an important component of research - it supports such things as participant confidentiality, organization, work efficiency, quality control, analysis….

As data are processed (cleaned) new 'versions' are created - from raw data, to the versions which will be used for analysis, and beyond.

Qualitative research example:

'Raw' audio data = the original digital audio recording

'Raw' transcript data = the original and unaltered transcript (text)

'Master' transcript data = the processed transcript → e.g., further de-identified, interviewer comments, typos fixed, etc.

'Analytic' transcript = working copy used for analysis/importing into analytic software

Case study project:

"Living with life-threatening illness: Narratives of liminality"

PI: Dr. Laurene Sheilds, University of Victoria; CIHR funded
Marcy Antonio, Research Coordinator

CONTEXT AND SETTING

Living with Serious Illness: Narratives of Liminality

Participants	Illness Groups	Sites
• Patients • Family Caregivers	• COPD • Cancer • Renal disease • Heart failure	• Edmonton • Victoria

- Principal and co-investigators (7)
- Interviewers (6)
- Transcribers (~5)

5

QUANTITY OF QUALITATIVE DATA

Living with Serious Illness: Narratives of Liminality

Data		Participants		
• Audio Interviews (3) • Field Notes (3+) • Transcripts (3) • Photographs (3+)	x	• Patients (n~40) • Family members (n~40)	=	960

ANALYSIS

9

ANALYSIS

Unit of Analysis	Analytic Frame	Sample
Case	Each participant (patient and family members) as a case including all fatal chronic conditions	n=80
Group	All patients (all conditions)	n=40
Sub-Group	Patients by condition (ESRD, HF, COPD, cancer)	n=4(10)
Group	All family members (all conditions)	n=40
Sub-Group	Family members by condition (ESRD, HF, COPD, cancer)	n=4(10)
Dyad Case	Each dyad (family member/patient) as a case (all conditions)	n=40
Group	All dyads patients/family member (all conditions)	n=80
Sub-Group	Dyads of patients/family members by condition	n=10(2)x4

10

Transcript File Naming Convention Table

This references the file names at the header on all of the transcripts.

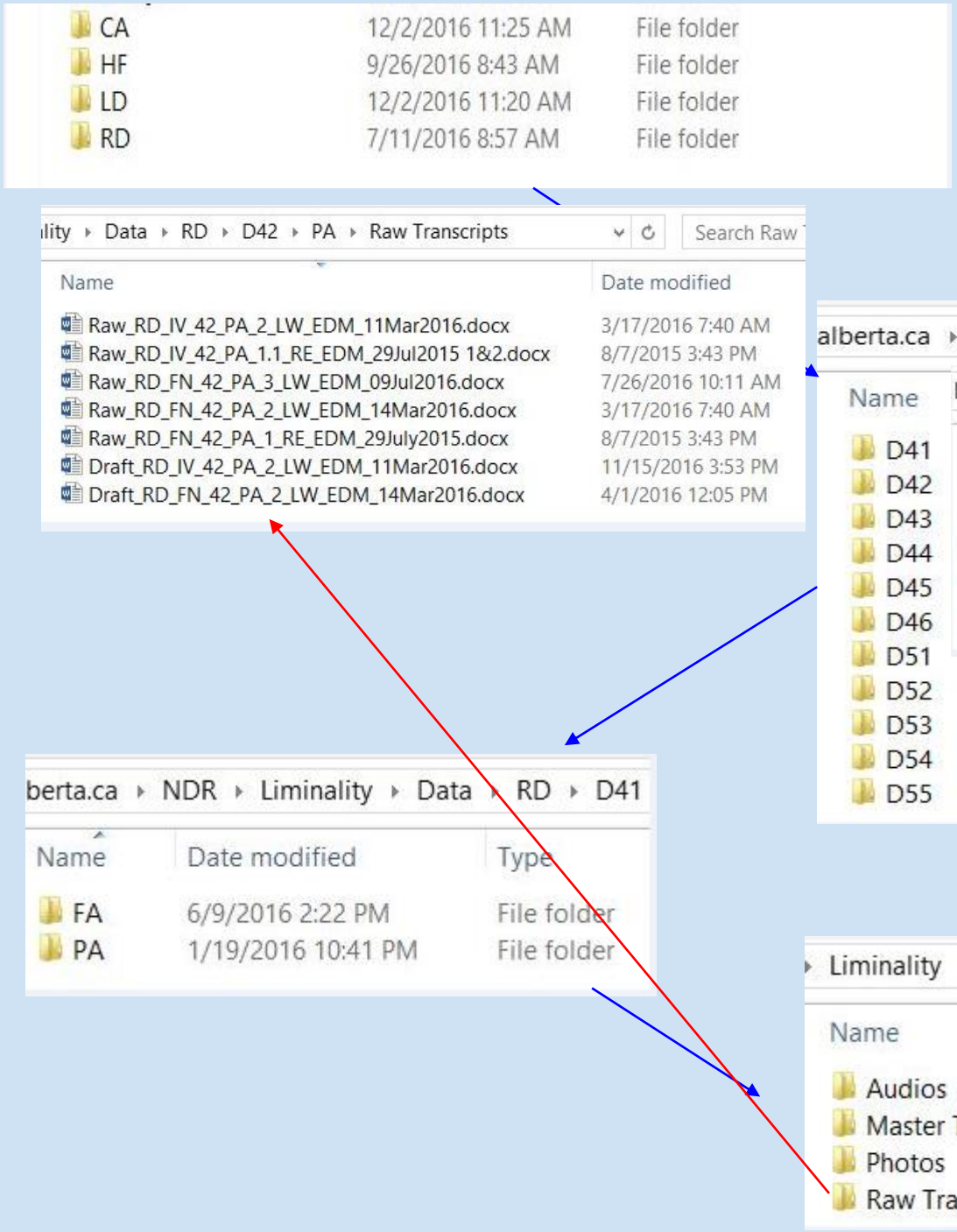
File Version *NOTE: This is only applicable for the Interviews (IN) and Transcripts (IVT)	**Illness**	**Data Type**	**Dyad Number** (Numerically as they enter the study; two digits)	**Dyad**	**Interview / FN/ Photo TYPE** *NOTE: Interviews with multiple parts (Part 1, 2, etc) use decimals (i.e., 1.1, 1.2)	**Interviewer** (first and last initial)	**Location** (first three letters of city)	**Day**	**Month**	**Year**
Raw	HF=Heart Failure	IV = Audio Interview	01-20=HF	PA=Patient	1= 1st in-person interview / 1st photo	Joanna=JC	EDM=Edmonton	Two digits	First three letters	Use four Digits
Master	LD = COPD/Lung Disease	IVT = Interview Transcribed	21-40=LD	FA=Family	2=2nd in-person interview / second photo	LW=Lacie	VIC = Victoria			
Draft (use when cleaning transcript)	RD=End Stage Renal Disease	FN=Field Notes	41-60=RD		3=Phone interview / third photo	LD= Lindsay				
	CA = Advanced Cancer	PH=Photo	61-80=CA		OTH#= Other (e.g. field note during intake call)	LC=Lynn				
						MA=Marcy				
						RE=Rebecca				

EXAMPLE 2: **MASTER_LD_IVT_39_PA_01_MA_VIC_15DEC2015**

Master Transcript	COPD	Interview	Dyad 01	Participant	First Interview	Marcy	Victoria	December 15, 2015

EXAMPLE 1: **Raw_CA_FN_01_FA_OTH1_LD_VIC_15Apr2016**

Raw Transcript	Advanced Cancer	Field note	Dyad 01	Family Member	Field note made at a time not specific to an interview (e.g. dropping off the camera)	Lindsay	Victoria	April 15, 2016


CA 12/2/2016 11:25 AM File folder
HF 9/26/2016 8:43 AM File folder
LD 12/2/2016 11:20 AM File folder
RD 7/11/2016 8:57 AM File folder
lity ▸ Data ▸ RD ▸ D42 ▸ PA ▸ Raw Transcripts
Search Raw
Name
Date modified
Raw_RD_IV_42_PA_2_LW_EDM_11Mar2016.docx 3/17/2016 7:40 AM
Raw_RD_IV_42_PA_1.1_RE_EDM_29Jul2015 1&2.docx 8/7/2015 3:43 PM
Raw_RD_FN_42_PA_3_LW_EDM_09Jul2016.docx 7/26/2016 10:11 AM
Raw_RD_FN_42_PA_2_LW_EDM_14Mar2016.docx 3/17/2016 7:40 AM
Raw_RD_FN_42_PA_1_RE_EDM_29July2015.docx 8/7/2015 3:43 PM
Draft_RD_IV_42_PA_2_LW_EDM_11Mar2016.docx 11/15/2016 3:53 PM
Draft_RD_FN_42_PA_2_LW_EDM_14Mar2016.docx 4/1/2016 12:05 PM
alberta.ca ▸ NDR ▸ Liminality ▸ Data ▸ RD
Name
lity ▸ Data ▸ RD ▸ D42 ▸ PA ▸ Master Transcripts
Name
Master_RD_FN_42_PA_1_RE_EDM_29July2015.docx
Master_RD_FN_42_PA_2_LW_EDM_14Mar2016.docx
Master_RD_IV_42_PA_2_LW_EDM_11Mar2016.docx
Master_RD_IV_FN_42_PA_3_LW_EDM_09Jul2016.docx
Master_RD_IVT_42_PA_1.1-1.2_RE_EDM_29Jul2015.docx
D41
D42
D43
D44
D45
D46
D51
D52 5/25/2016 2:45 PM File folder
D53 3/3/2016 12:54 PM File folder
D54 12/5/2016 5:30 PM File folder
D55 10/28/2016 4:01 PM File folder
berta.ca ▸ NDR ▸ Liminality ▸ Data ▸ RD ▸ D41
Name
Date modified
Type
FA 6/9/2016 2:22 PM File folder
PA 1/19/2016 10:41 PM File folder
Liminality ▸ Data ▸ RD ▸ D41 ▸ PA
Name
Date modified
Type
Audios 12/8/2016 2:11 PM File folder
Master Transcripts 12/8/2016 2:16 PM File folder
Photos 1/21/2016 8:06 AM File folder
Raw Transcripts 12/8/2016 2:06 PM File folder

3. Post-Research - **Publication requirements**

- Increasingly, journals are requiring research data to be made openly accessible
- Talk to researchers early about journals they may be interested in submitting to
- Find out what the journals' data policies are
- Most journals requiring data to be made openly available will have exceptions for data with legal and/or ethical considerations
- Refusal to share data are grounds for rejection

EXAMPLE: PLOS ONE

PLOS One is a peer-reviewed open access scientific journal published by the Public Library of Science (**PLOS**) since 2006. The journal covers primary research from any discipline within science and medicine.

3. Post-Research

Data deposit

We will be covering data deposit in-depth this afternoon!

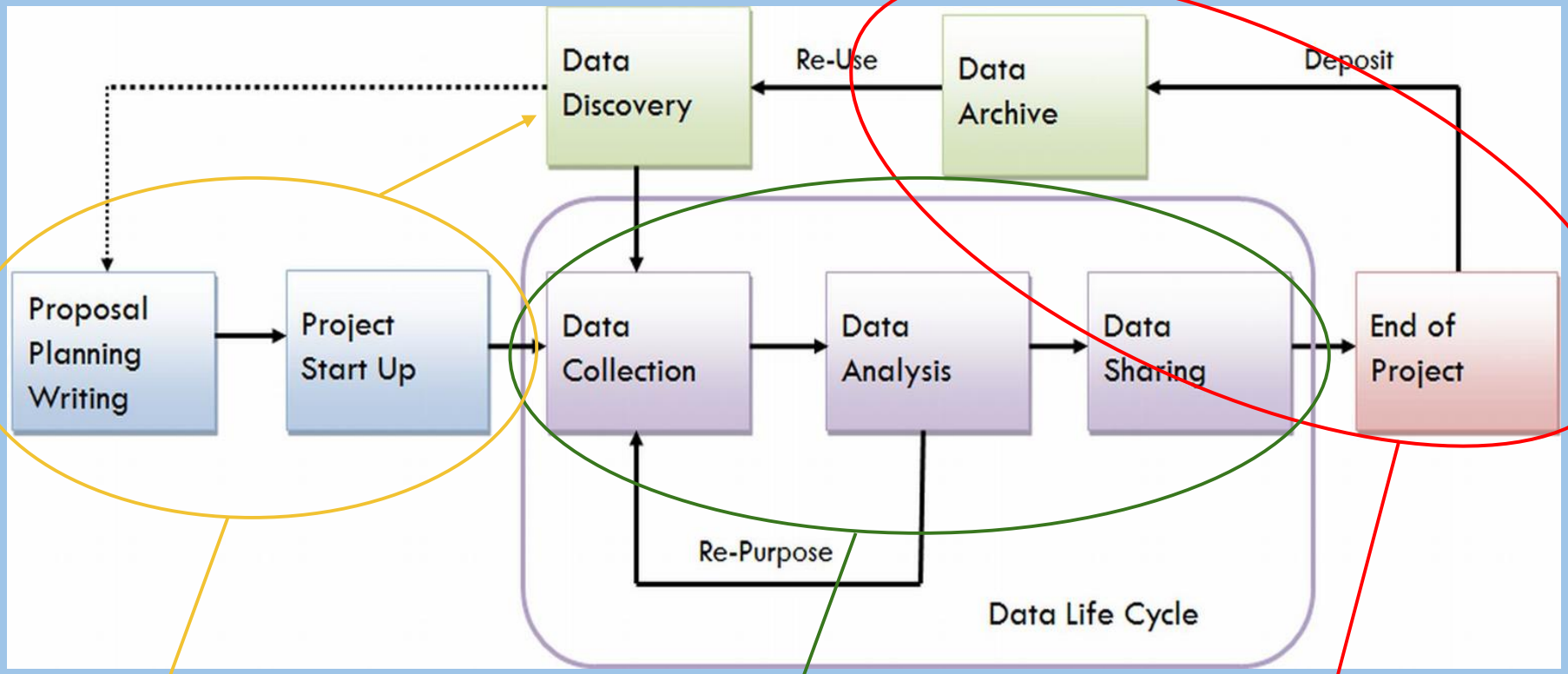
3. Post-Research

Data deposit

Tardis & Dr. Who Libraries:
https://www.pinterest.ca/ltlfreelibrary/tardis-and-dr-who-libraries/

RDM 'Best Practices': A bird's eye view

1. Overarching principles: FAIR; CARE; OCAP



2. Pre-Research:
   - Including RDM into funding applications
   - Funder requirements
   - Data Management Plans
   - Data discovery
   - Participant consent & Information Letters
   - Ethics applications

3. During Research:
   - Primary data collection
   - Data storage
   - Data transferring
   - Data access
   - Data documentation/metadata
   - File naming conventions

4. Post-Research:
   - Publication requirements
   - Data deposit

Questions & Discussion

James Doiron
RDM Services Coordinator, UofA Libraries
Academic Director, UofA Research Data Centre
Co-Chair, Portage DMP Expert Group

jdoiron@ualberta.ca

Shifting horizons slide