UNIVERSITY OF CALGARY

Data Management Planning

John Brosz, PhD
Coordinator, Research Data & Vis.

Kathryn Ruddock, MLIS
Manager, Digitization & Repository Services

Slides: http://dx.doi.org/10.5072/PRISM/31903

Outline

- What is research data management?
- Data Management & the granting agencies
- Data Management Planning
- Scenarios

Slides: http://dx.doi.org/10.5072/PRISM/31903

Research Data?

- Research data are pieces of (digital) information, structured through methodology for the purpose of producing new knowledge.

- A better term is research material and can include:
  - Images
  - Email, chat logs
  - Sound or video recordings
  - Presentations
  - Document data
  - Oral history & interviews
  - News stories
  - Bibliography
  - Measurements & statistics
  - 3D models, virtual environments, simulations
  - Software, source code
  - Transcripts
  - Procedures & workflows

Research Data Management

- Data Curation, Data Stewardship
- Application of standard practices towards the creation and analysis of data for current and future purposes

Why Research Data Management

- Save your time
- Preserve your data
- Maintain data integrity (versions)
- Meet requirements
- Easier collaboration
- Promote new discoveries (get more citations)
- Provide evidence
  - Increase reproducibility & verifiability

Research Data Management Stages

Before

- Using other's data
  - Finding data
  - Copyright
  - Data licensing
- Funding guidelines
- Data management plan
- Ethics

During

- Documentation & metadata
- Storage & data protection
- Managing sensitive data
- Collaboration

After

- Sharing data
  - Data repositories:
    - Local
    - Disciplinary
    - General
- Preservation / Archiving

Background – International Context

Background:

- Growing number of Research Data Management and sharing policies worldwide.
- Objectives:
  - Improve efficiency of research
  - Support re-use of data for new insights and discoveries
  - Foster collaboration
  - Facilitate greater transparency
- Trend is progressing in parallel with the movement towards open access publications

Source: *Comprehensive Brief on Research Data Management Policies*, Kathleen Shearer, 2015

Background – International Context

- In the US, Data management policies and requirements for DMP have been introduced by NIH (2001, 2003), NSF (2011), and more to come for federally funded agencies.
- In the UK, “Common Principles on Data Policy” issued by Research Council UK (2011), and most UK funders now require a DMP
- In the EU, a pilot action on data management requirements within the Horizon 2020 framework. The approach is “as open as possible, as closed as necessary.”

Source: *Comprehensive Brief on Research Data Management Policies*, Kathleen Shearer, 2015

Tri-Agency Open Access Policy on Publications (2015)

Publication-related Research Data

- **For recipients of CIHR funding:**
  - Deposit bioinformatics, atomic, and molecular coordinate data into the appropriate public database immediately upon publication of research results.
  - Retain original data sets for a minimum of five years after the end of the grant (published or not).

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

Tri-Agency Statement of Principles on Digital Data Management

Rationale and benefits:

- CIHR, NSERC and SSHRC are publically funded federal granting agencies
- Desire to ensure widest possible dissemination of publicly-funded research results in order to:
  - Advance knowledge
  - Avoid research duplication and encourage re-use
  - Maximize benefits to Canadians
  - Showcase accomplishments of Canadian researchers
- Statement of Principles on Digital Data Management (released June 2016)

Tri-Agency Statement of Principles on Digital Data Management

Statement expectations

- **Data Management Planning** is necessary at all stages of the research project lifecycle
- Research data must be managed in agreement with all **commercial, legal and ethical obligations**.
- Data should be managed in accordance with **relevant standards and best practices**
- All research data should be accompanied by **metadata**

Tri-Agency Statement of Principles on Digital Data Management

Statement expectations

- Research data resulting from agency funding should normally be **preserved** in a publicly accessible, secure and curated repository for **discovery** and **reuse**.
- Data should be shared as **early as possible** in the research process
- Data are significant and legitimate products of research and should be **acknowledged and cited** as such.
- Data Management should be **efficient and cost-effective**

Tri-Agency Statement of Principles on Digital Data Management

Statement expectations

**"All data need to be managed,
but not all data need to be shared or preserved"**

- Consider costs, benefits, and legal or ethical obligations.
  - Consider whether stored information would be useful for secondary analysis, replication & verification, etc.
  - Ensure confidentiality, privacy and legal obligations are respected.
  - Consider whether data need to be de-identified or made available with restricted access.
- Rationale for data preservation, sharing and retention is defined in DMP.

What to Expect in the Future?

- The release of the statement represents *one step* in the process of reviewing and enhancing the data management requirements for Tri-Agency supported research.
- SSHRC has completed a pilot project with select awardees (retroactive data management plan exercise) to guide future directions in policy
- Intend to consult further with stakeholders
- Policy development

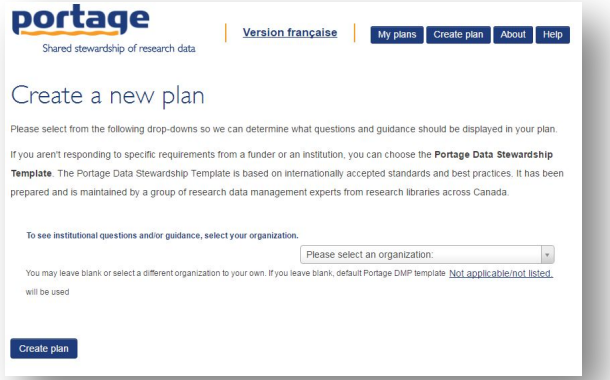
Data Management Plans (DMPs)

- Short, living document
- Describe how you will
  organize, store, and share
  your research data through the lifetime of the research
- Does not contain data

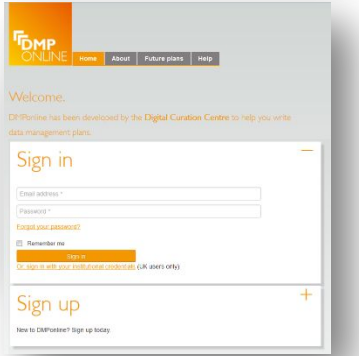
Creating a DMP

How to create a plan?

- DMP Assistant (Canadian, bilingual)
  - https://assistant.portagenetwork.ca
  - Asks questions, your answers form a DMP
  - Templates – Portage Template



- DMP Online (UK)
  - https://dmponline.dcc.ac.uk/
- DMPTool (US)
  - https://dmptool.org/



portage

Shared stewardship of research data

Version française

Home About Help

https://assistant.portagenetwork.ca

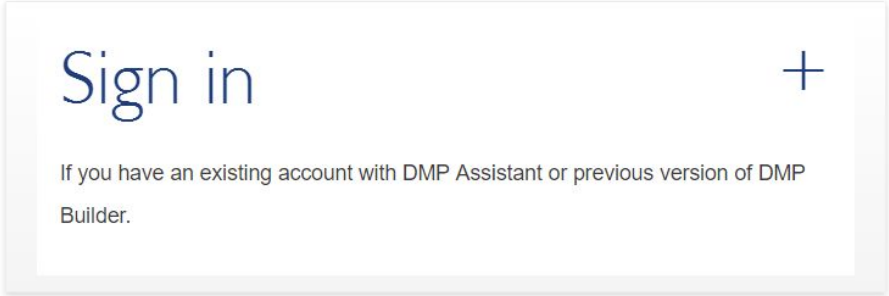
DMP Assistant is a bilingual tool for preparing data management plans (DMPs). The tool follows best practices in data stewardship and walks researchers step-by-step through key questions about data management.

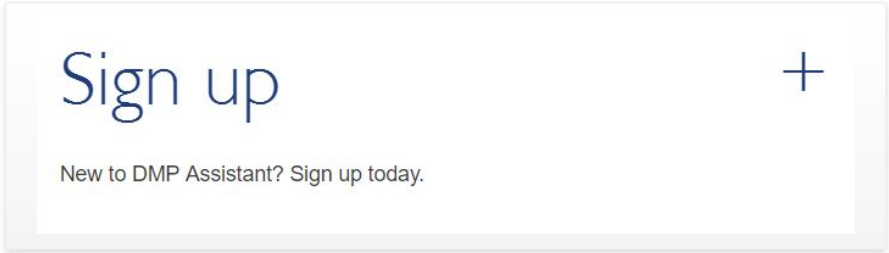
Step 1 Sign up with DMP Assistant

Step 2 Sign in and select a template under Organizations. The Portage template is the default.

Step 3 Answer the questions that are relevant to your work. Guidance and examples are provided.

Step 4 Revisit the tool throughout your research to review or revise your answers.





Please note that we are currently working on single sign-in authentication. For now, please create a new DMP Assistant account. You will have the option to link your DMP Assistant account to your campus ID when that feature becomes available.

Create a Plan

Create a new plan

Please select from the following drop-downs so we can determine what questions and guidance should be displayed in your plan.

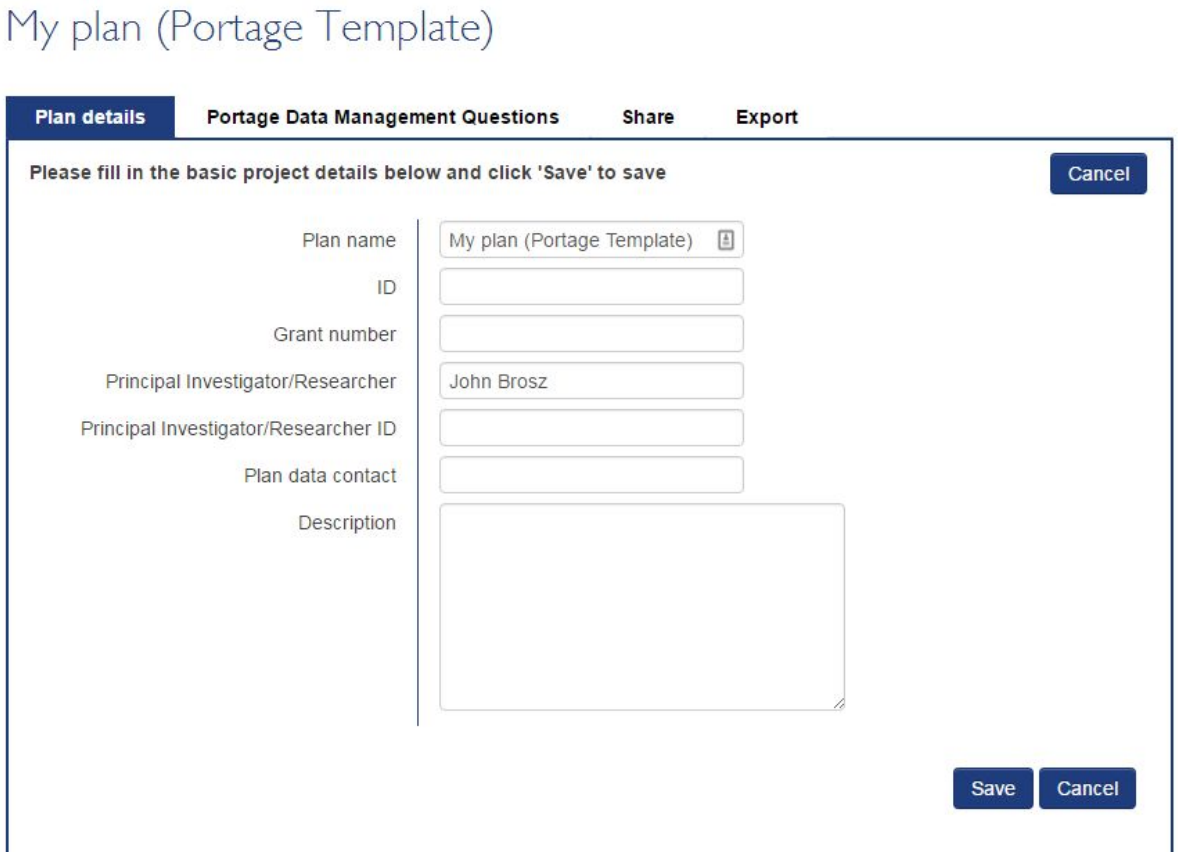
If you aren't responding to specific requirements from a funder or an institution, you can choose the **Portage Data Stewardship Template**. The Portage Data Stewardship Template is based on internationally accepted standards and best practices. It has been prepared and is maintained by a group of research data management experts from research libraries across Canada.

To see institutional questions and/or guidance, select your organization. University of Calgary

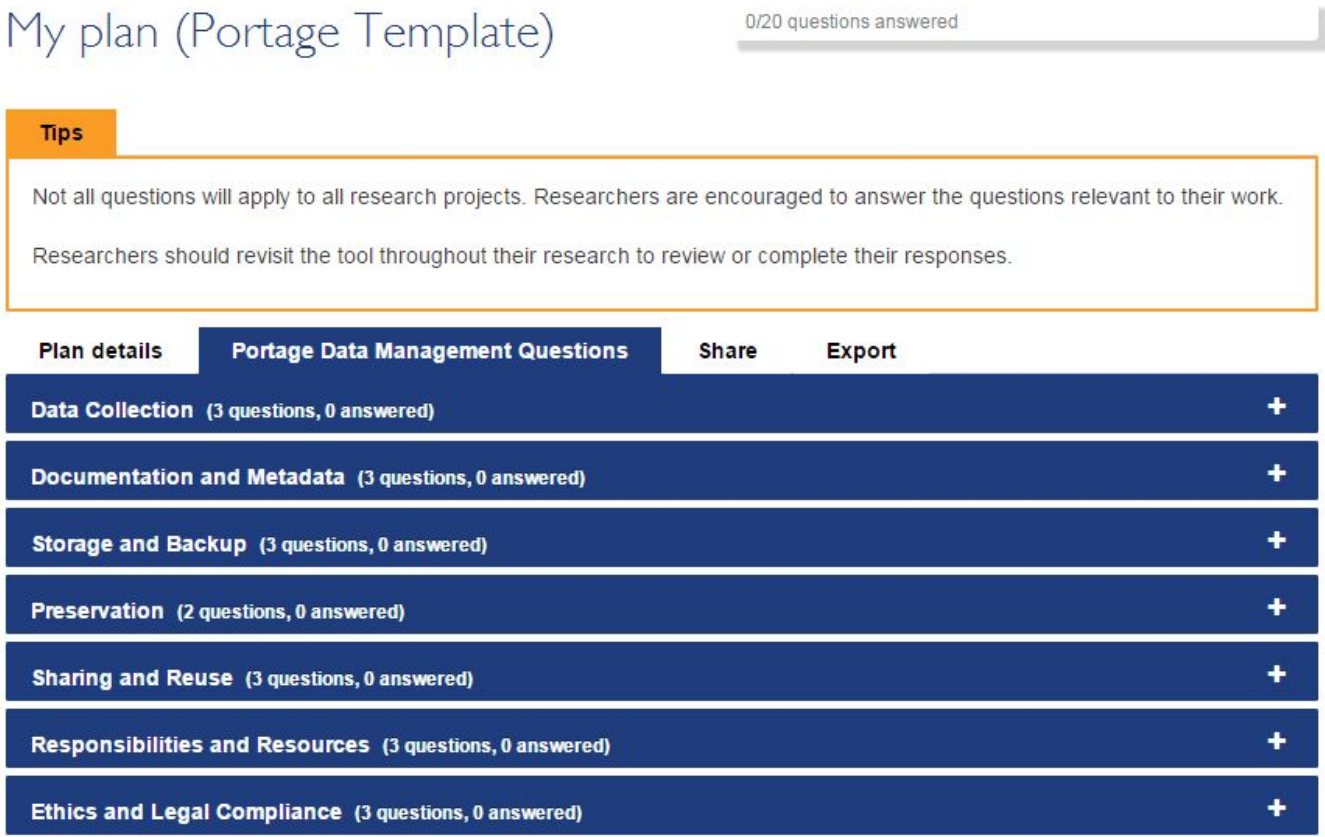
You may leave blank or select a different organization to your own. If you leave blank, default Portage DMP template will be used Not applicable/not listed.

Create plan

Metadata


My plan (Portage Template)
Plan details
Portage Data Management Questions
Share
Export
Please fill in the basic project details below and click 'Save' to save
Cancel
Plan name
My plan (Portage Template)
ID
Grant number
Principal Investigator/Researcher
John Brosz
Principal Investigator/Researcher ID
Plan data contact
Description
Save
Cancel

20 Questions



Questions

Identify who will be responsible for managing this project's data during and after the project and the major data management tasks for which they will be responsible.

Save

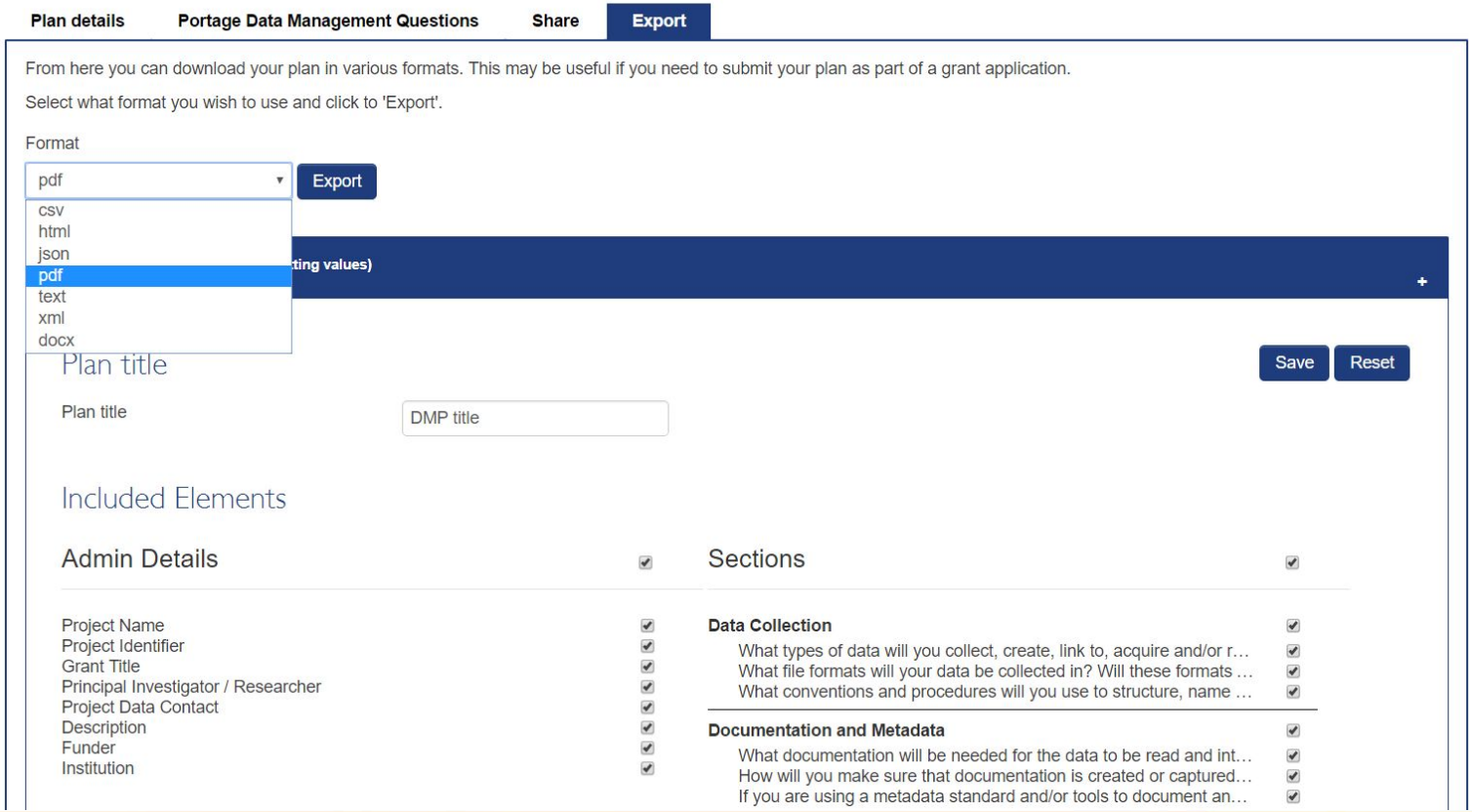
Guidance | Share note

Portage Guidance +

Your data management plan has identified important data activities in your project. Identify who will be responsible -- individuals or organizations -- for carrying out these parts of your data management plan. This could also include the timeframe associated with these staff responsibilities and any training needed to prepare staff for these duties.

Export

My plan (Portage Template)



1. Data collection
2. Documentation & metadata
3. Storage & Backup
4. Preservation
5. Sharing & Reuse
6. Responsibility & Resources
7. Ethics & Legal Compliance

1. Data collection
   - What data are you creating/acquiring?
   - File types
   - File naming conventions
2. Documentation & metadata
3. Storage & Backup
4. Preservation
5. Sharing & Reuse
6. Responsibility & Resources
7. Ethics & Legal Compliance

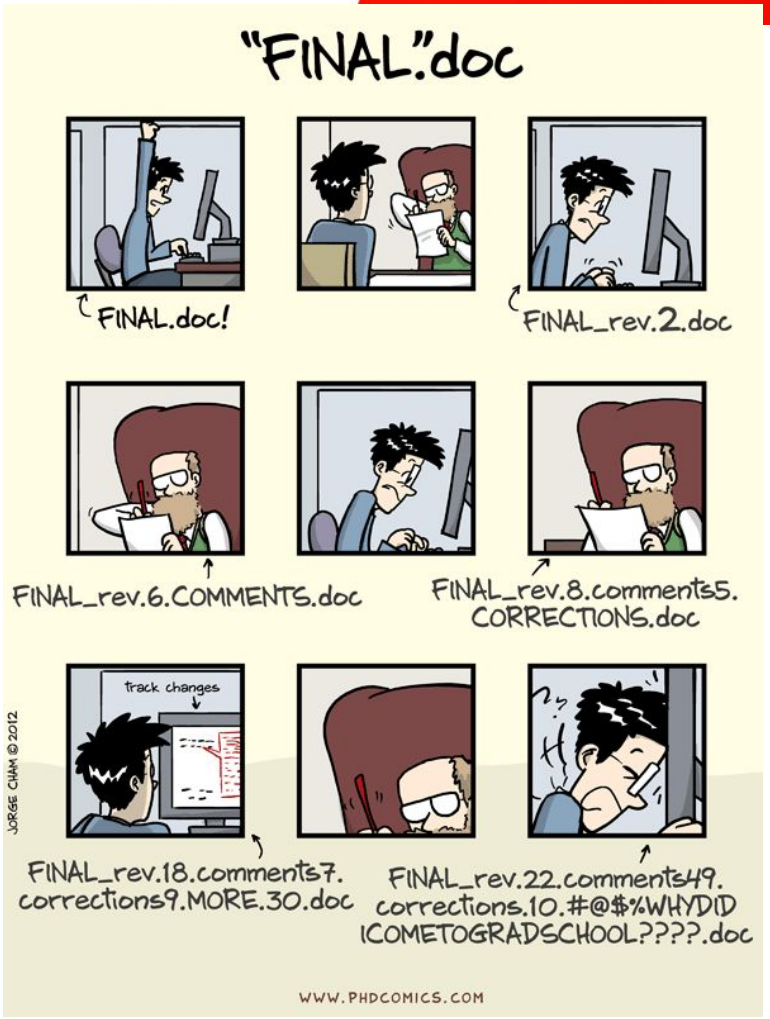
1. Data collection
   - What data are you creating/acquiring?
   - File types
   - File naming conventions
2. Documentation & metadata
3. Storage & Backup
4. Preservation
5. Sharing & Reuse
6. Responsibility & Resources
7. Ethics & Legal Compliance

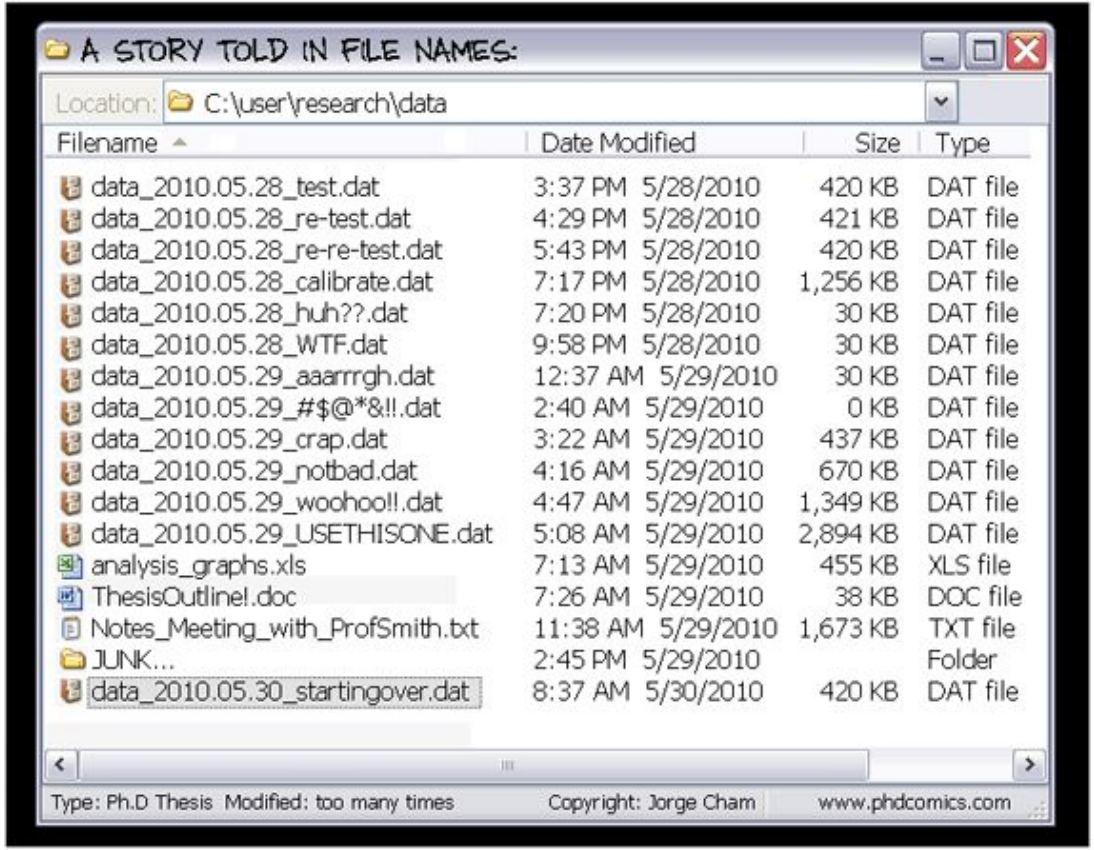
File Formats

- If at all possible, don't use proprietary file formats
- If not possible, try to also include open versions (.rtf with .docx)

Recommended File Formats:	
Images: JPG, PNG, PDF, TIFF, BMP	**Sound**: MP3, FLAC
Spreadsheets: CSV	**Text**: TXT, CSV, PDF/A, ASCII, UTF-8
Video: MPG, MOV, AVI	**Ebooks**: EPUB
Databases: XML, CSV	

1. Data collection
   - What data are you creating/acquiring?
   - File types
   - File naming conventions
2. Documentation & metadata
3. Storage & Backup
4. Preservation
5. Sharing & Reuse
6. Responsibility & Resources
7. Ethics & Legal Compliance





http://phdcomics.com/comics.php?f=1531
http://phdcomics.com/comics.php?f=1323

File naming and organizing

- Can someone else understand/use your data files?
  - Now?
  - Tomorrow?
  - In 5 years?!?
- Document **everything** you do with file naming, changes to files, transformations to data, etc.
- Some extra work in the beginning and as you go will save you **hours** of time later, and possibly your entire research project!

File naming conventions make life easier!

Naming conventions should be:

- **Descriptive**
- **Consistent**

Consider Including:

- Unique identifier (Project Name or Grant # as folder)
- Project or research data name
- Conditions (Lab instrument, Solvent, Temperature, etc.)
- Run of experiment (sequential)
- Date (in file properties too)
- Version #

Files **with** a naming convention:

- 20130503_DOEProject_DesignDocument_Smith_v2-01.docx
- 20130709_DOEProject_MasterData_Jones_v1-00.xlsx
- 20130825_DOEProject_Ex1Test1_Data_Gonzalez_v3-03.xlsx
- 20130825_DOEProject_Ex1Test1_Documentation_Gonzalez_v3-03.xlsx
- 20131002_DOEProject_Ex1Test2_Data_Gonzalez_v1-01.xlsx
- 20141023_DOEProject_ProjectMeetingNotes_Kramer_v1-00.docx

1. Data collection
2. Documentation & metadata
   - What documentation is needed to work with your data?
   - How will you ensure documentation created?
   - Describe metadata & associated tools
3. Storage & Backup
4. Preservation
5. Sharing & Reuse
6. Responsibility & Resources
7. Ethics & Legal Compliance

1. Data collection
2. Documentation & metadata
   - What documentation is needed to work with your data?
   - How will you ensure documentation created?
   - Describe metadata & associated tools
3. Storage & Backup
4. Preservation
5. Sharing & Reuse
6. Responsibility & Resources
7. Ethics & Legal Compliance

1. Data collection
2. Documentation & metadata
   - What documentation is needed to work with your data?
   - How will you ensure documentation created?
   - Describe metadata & associated tools
3. Storage & Backup
4. Preservation
5. Sharing & Reuse
6. Responsibility & Resources
7. Ethics & Legal Compliance

Metadata

- Data *about* your data

- The main purpose of metadata is to facilitate in the discovery of relevant information. Metadata assists in resource discovery by "allowing resources to be found by relevant criteria, identifying resources, bringing similar resources together, distinguishing dissimilar resources, and giving location information.”

https://www.wikiwand.com/en/Metadata

- Everything you’d need to recreate your data

UNIVERSITY OF CALGARY

No Metadata

Name	Artist	Album by Artist	Time	Track #	Rating	Plays	Last Played	Date Added
			5:20	18				4/20/04 7:14 AM
Track 18			4:24	18 of 19				11/26/07 10:35 PM
You've Lost that Lovin Feeeeeeling			3:08	17 of 20				1/3/04 11:18 AM
Track 17			3:27	17 of 19				11/26/07 10:34 PM
Track 17			3:19	17 of 20		1	3/21/03 10:50 PM	5/2/02 2:04 PM
			5:38	16				4/20/04 7:14 AM
Track 16			3:26	16 of 19				11/26/07 10:34 PM
Track 16			3:50	16 of 18				5/10/03 5:35 PM
Beth			2:46	16 of 20				5/2/02 2:04 PM
			4:27	15				4/20/04 7:14 AM
Track 15			4:33	15 of 19				11/26/07 10:34 PM
Track 15			2:36	15 of 18				5/10/03 5:35 PM
			4:28	14				4/20/04 7:14 AM
Track 14			6:37	14 of 19				11/26/07 10:33 PM
Track 14			3:02	14 of 18				5/10/03 5:35 PM
Track 14			4:52	14 of 14		1	3/21/03 11:15 PM	5/2/02 2:04 PM
Track 14			3:06	14 of 14		1	3/21/03 11:21 PM	5/2/02 2:03 PM
mystery5			2:44	14 of 24				8/12/04 11:33 PM
			4:48	13				4/20/04 7:14 AM
Track 13			4:13	13 of 19				11/26/07 10:33 PM
Track 13			1:12	13 of 21				10/8/03 1:35 AM
Track 13			4:45	13 of 18				5/10/03 5:35 PM
Track 13			3:15	13 of 14		1	3/21/03 11:26 PM	5/2/02 2:04 PM
Detroit Rock City			3:35	13 of 20				5/2/02 2:04 PM
			4:14	12				4/20/04 7:14 AM
Track 08			4:31	8 of 19				11/26/07 10:32 PM
Track 08			3:19	8 of 10				6/6/05 4:58 AM
Track 08			3:32	8 of 19				5/2/02 2:04 PM
Track 08			3:38	8 of 14		1	7/7/03 6:04 PM	5/2/02 2:04 PM
Track 08			4:27	8 of 8				5/2/02 2:03 PM
Sailors Bid			0:35	8 of 11				3/12/09 12:32 PM
Mystery Song 1			3:52	8				6/2/05 9:00 PM
			4:08	7				4/20/04 7:14 AM
07 07 07 07 07 07 CD 2 07 When I think of you 1			10:19	7				4/20/04 7:14 AM
Track 07			8:17	7 of 19				11/26/07 10:31 PM
Track 07			8:17	7 of 19				10/24/06 11:55 PM

??!?!??

https://flic.kr/p/4toCrw

Metadata

What is Exif data?

Exif data is a record of the settings a camera used to take a photo or video. This information is embedded into the files the camera saves, and we read and display it here.

Dates

Taken on	April 24, 2012 at 12.57PM MDT
Posted to Flickr	April 24, 2012 at 1.13PM MDT

Exif data

Camera	Nikon D40
Exposure	0.005 sec (1/200)
Aperture	f/7.1
Focal Length	24 mm
Focal Length	23.8 mm
ISO Speed	200
Exposure Bias	0 EV
Flash	No Flash
Orientation	Horizontal (normal)
X-Resolution	300 dpi
Y-Resolution	300 dpi
Software	Ver.1.10
Date and Time (Modified)	2012:04:24 12:57:32
YCbCr Positioning	Co-sited

UNIVERSITY OF CALGARY

Metadata Formats

- Documentation for understanding & re-use
  - Readme File
  - Data Dictionary
  - Codebook
- Structured documentation in XML
  - DDI
  - FGDC

UNIVERSITY OF CALGARY

Metadata in Research

Project Documentation	Dataset Documentation
• Context of data collection • Data collection methods • Structure, organization of data files • Data sources used • Data validation, quality assurance • Transformations of data from the raw data through analysis • Information on confidentiality, access and use conditions	• Variable names and descriptions • Units of measure • Instruments • Explanation of codes and schemas used • Algorithms used to transform data • File format and software (version) used

Metadata

- Describe your data as thoroughly as possible
- Use standards as appropriate
  - DDI - Archiving and Social Science
  - Darwin Core – Biology, Archaeology
  - DIF - Scientific data sets
  - CSDGM - Geographic data
  - TEI - Humanities, social sciences and linguistics
  - Etc…
- Most Data Repositories will walk you through minimum metadata, but the more, the better.
- Working backwards from where you want your data to end up, check the repository

1. Data collection
2. Documentation & metadata
3. Storage & Backup
   - How much space & for how long?
   - Team access
   - Backup?
4. Preservation
5. Sharing & Reuse
6. Responsibility & Resources
7. Ethics & Legal Compliance

Backup

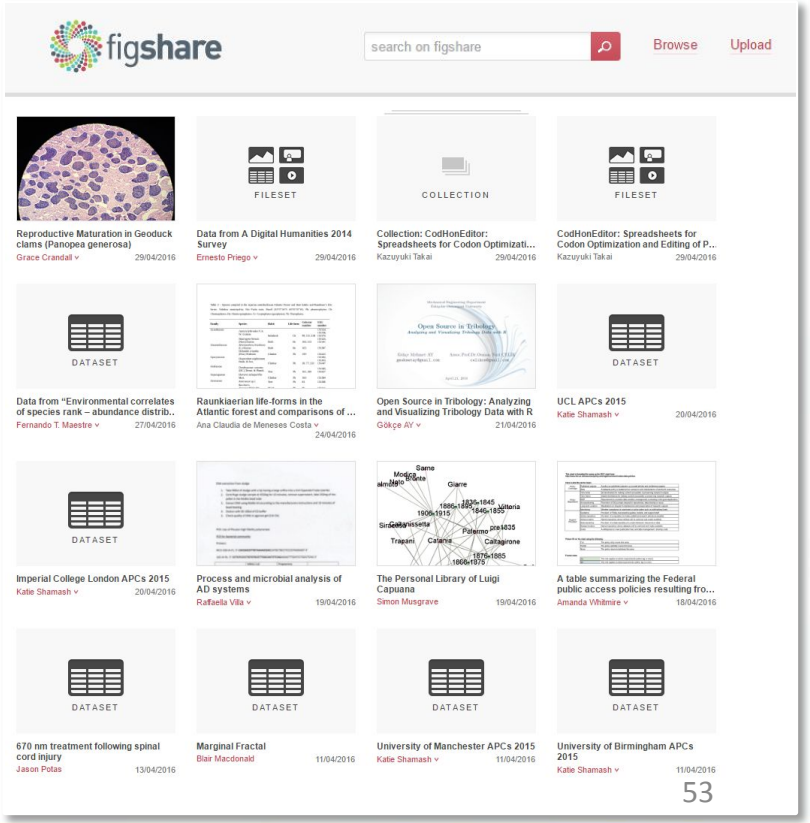
- **3 – 2 – 1 Rule**
  - Make **3** copies
  - On **2** different media
  - And **1** should be “offsite”
- Keep your anti-virus software up-to-date
- Keep your password**s** secure
- Consider using encryption
- Do not store master data copies on personal computers or laptops

1. Data collection
2. Documentation & metadata
3. Storage & Backup
4. Preservation
   - What happens after the project finishes?
   - Digital obsolescence
5. Sharing & Reuse
6. Responsibility & Resources
7. Ethics & Legal Compliance

1. Data collection
2. Documentation & metadata
3. Storage & Backup
4. Preservation
5. Sharing & Reuse
   - Will your data be shared?
   - End user license?
   - How are you sharing?
6. Responsibility & Resources
7. Ethics & Legal Compliance

Sharing: Data Repositories

- Designed for data
- Securely-stored
- Authoritative copy of your data
  - Provides with a DOI (or similar)
  - Easily citable
- Searched index
  - Discoverable through google, etc.
- Open access
- Some are free, others are not

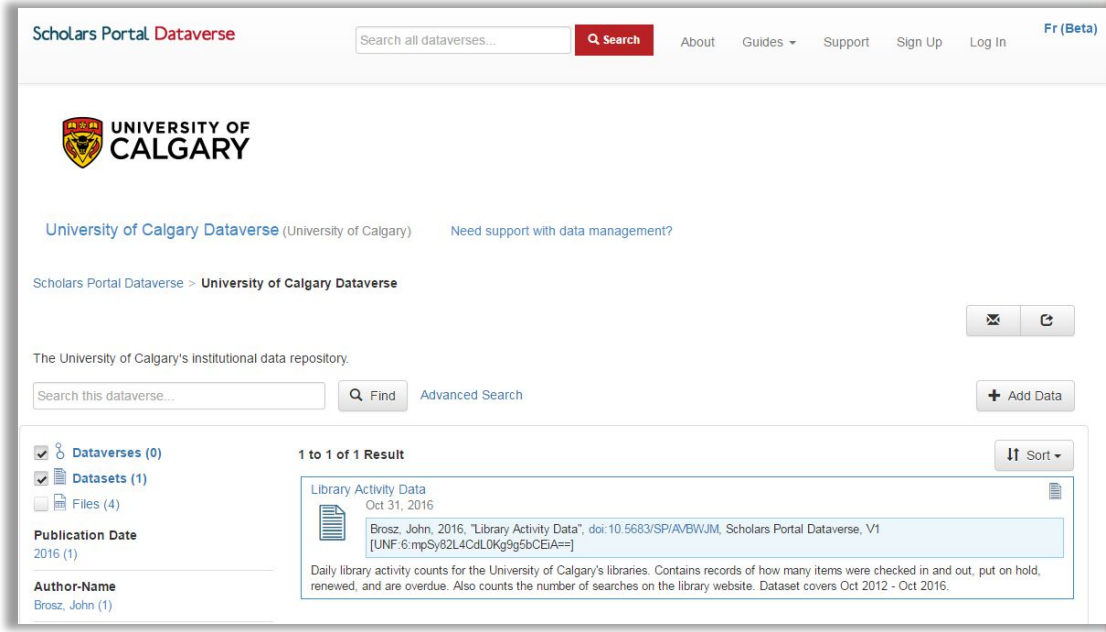


Data Repositories

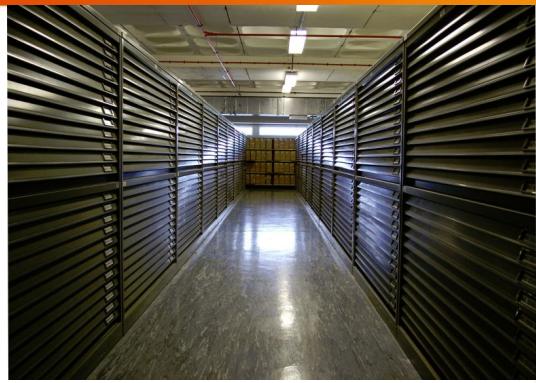
University of Calgary Data Repository

https://dataverse.scholarsportal.info/dataverse/calgary

- Free
- File versioning
- Dataset specific metadata fields
- Access control
- DOI
- Multiple files, up to 2GB each



Data Repositories



- Disciplinary
  - GenBank – molecular biology & genetics
  - tDAR – archaeology
  - NIH repositories list – variety of health sciences data repositories
  - Virtual Astronomical Observatory – astronomical telescope data
  - Cambridge Structural Database – molecular crystal structures
  - GitHub – source code
  - PANGEA – life and earth sciences
  - Good list: https://www.lib.umn.edu/datamanagement/datacenters
- General
  - ICPSR – social science data
  - FigShare
  - Zenodo
  - Dryad Digital Repository (DataDryad)

Research Data Repository Directory

- Re3Data.org

1. Data collection
2. Documentation & metadata
3. Storage & Backup
4. Preservation
5. Sharing & Reuse
6. Responsibility & Resources
   - Who does what?
   - What happens if . . .
   - Costs
7. Ethics & Legal Compliance

1. Data collection
2. Documentation & metadata
3. Storage & Backup
4. Preservation
5. Sharing & Reuse
6. Responsibility & Resources
7. Ethics & Legal Compliance
   - Security for sensitive data
   - Legal, ethic, & IP issues

Sharing Concerns

- You **must** abide by your ethics agreement & informed consent.
- To ensure the confidentiality of your subjects watch for:
  - Personally Identifiable Information
  - Protected Health Information
  - Sensitive Information
- Watch for info that seems generic, but still might resolve as an individual:
  - Race or gender
  - Geography
  - Specific age
- What if you can't scrub your data?
  - Embargoes
  - Technological Access Restrictions
  - Data Use Agreements

UNIVERSITY OF CALGARY

Licenses

- Creative Commons 0 (CC0) is recommended for research data
  - E.g., freely available without restriction
- Why not “CC BY” or others?
  - Data (facts) are not copyrightable but works of authorship are . . . this makes this a messy, complicated issue.
  - If you are not sure copyright, people wanting to use your data will not be either.
  - CC0 makes it more likely your data will be reused.
  - Scholars already have very good reasons to provide attribution.

https://datapub.cdlib.org/2016/09/19/cc-by-and-data-not-always-a-good-fit

UNIVERSITY OF CALGARY

Very Brief Guide

1. Save your raw data
2. Backup your data
3. Describe your data
4. Process your data
5. Archive & preserve your data

From http://researchdata.library.ubc.ca/files/files/2017/07/GoodEnoughResearchDataManagement_V1.1_20170705.pdf

UNIVERSITY OF CALGARY

Very Brief Guide

1. Save your raw data
   - Don't overwrite original data with cleaned version.
   - Protect your original data (master file) by locking it or making it read only.
   - Refer to this master file if things go wrong.

From http://researchdata.library.ubc.ca/files/files/2017/07/GoodEnoughResearchDataManagement_V1.1_20170705.pdf

UNIVERSITY OF CALGARY

Very Brief Guide

1. Save your raw data
2. Backup your data
3. Describe your data
   - Machine Friendly. Describe your dataset with a metadata standard (DataCite, Dublin Core, DDI)
   - Human Friendly. Describe your variables to colleagues will understand what they are. Use descriptive, clear variables names.
   - Use "NA" for missing data, blank cells confuse everyone.
   - Convert to non-proprietary formats
   - Use a file naming standard.

From http://researchdata.library.ubc.ca/files/files/2017/07/GoodEnoughResearchDataManagement_V1.1_20170705.pdf

UNIVERSITY OF CALGARY

Very Brief Guide

1. Save your raw data
2. Backup your data
3. Describe your data
4. Process your data
   - Make each column a variable
   - Make each row an observation
   - Store units of measure as metadata in their own column
   - Document each step of processing in a README file (or similar).

From http://researchdata.library.ubc.ca/files/files/2017/07/GoodEnoughResearchDataManagement_V1.1_20170705.pdf

UNIVERSITY OF CALGARY

Very Brief Guide

1. Save your raw data
2. Backup your data
3. Describe your data
4. Process your data
5. Archive & preserve your data
   - Submit final data files (where possible) to a repository with a DOI
   - Provide good metadata for your study so other people can find it.

From http://researchdata.library.ubc.ca/files/files/2017/07/GoodEnoughResearchDataManagement_V1.1_20170705.pdf

UNIVERSITY OF CALGARY

Scenarios

1. Medicine
2. Business
3. Arts
4. Engineering

https://assistant.portagenetwork.ca

Scenario 1

Medicine

- Louise is a new PhD student in the School of Public Health and Preventive Medicine. Her PhD topic relates to policy interventions to prevent the outbreak of infectious diseases like bird flu. She is interested in this topic because of her work as a policy analyst with AHS and her background in volunteering in developing countries, and sees completing the PhD as a good way to further her policy career as well as her interests in social development.
- Louise's research will involve a number of field interviews with health workers and policy makers in Canada, Vietnam, Indonesia and China. She has an iPod and thought that she would use this to make audio recordings of the interviews, which she will later analyze (possibly using NVivo).
- Louise also wants to access the policy documents of government agencies and health service providers (including hospitals) in Calgary and other jurisdictions in Canada and overseas. She thinks she will do some kind of content analysis on these, probably also using NVivo. Some agencies freely provide these documents on their websites, while other agencies have internal documents that are not readily available to the general public, which she may have to approach the organizations for directly.
- Louise wants to test her hypothesis that a speedy response from policy makers can reduce the spread of infectious diseases. This will require doing some cross-analysis of her findings from the policy documentation and interviews along with the World Health Organisation's *Cumulative number of confirmed human cases of avian influenza A(H5N1)* dataset, which is available for download from the WHO website as a series of PDFs published monthly.
- In doing her literature review there are a number of industry publications and academic journals that Louise has identified as potential places in which she might try to publish later. There are also some big international conferences coming up, and her supervisor has encouraged her to consider presenting her results at these.

UNIVERSITY OF CALGARY

Scenario 2

Business

- Gemma is about to start a PhD in the Haskayne Faculty of Business. Gemma worked as a stockbroker in London for several years, but is increasingly interested in environmental issues. For her PhD, she wants to track the relative success of shares included in ‘ethical investment' portfolios, compared to more general investments. She also wants to look at the newspaper coverage given to ethical investment in the financial sections of major Canadian newspapers to see if it has grown at the same rate as the number of ethical products in the market has grown.
- Gemma has already discovered that she can access TSX information through a database hosted by a not-for-profit company. This data goes back to 1991 but the most recent results can take several months to appear. The data is accessed via a web interface and the results that Gemma receives from her searches (which have a certain number of parameters) are put up on a server from where she can download them as a .csv file. The files only stay on Sirca's server for a month - after that time they are deleted.
- Gemma thinks she will probably only need Excel to do her analysis on the stock data - she has a copy of Microsoft Office installed on her laptop and plans to continue using this software.
- Gemma thinks that the best way to investigate the newspaper coverage would be to download the full text of lots of newspaper articles from the Library's databases and then load these into a software program called Leximancer, which is designed for textual analysis of the kind she wants to do. This tool was developed by academic researchers but has since been spun out into a small company. Gemma asked Ucalgary IT about Leximancer but they said the tool is not supported because there are only a few users of it locally. Nevertheless, a friend of Gemma's has found it so useful that he is paying the monthly subscription out of his own pocket and has recommended that Gemma do the same.
- Gemma thought her project was going really well, but her supervisor recently suggested that it might be better if she focused on more than one national market, and has suggested that she should think about including other countries such as the United States and Australia as part of her study.

Scenario 3

Arts

- Lachlan has recently started a PhD in Performance Studies. He is interested in the history of circus arts in North America, and developed this interest while doing paid and voluntary work as an arts administrator.
- Lachlan will be doing archival research in state and city archives in Calgary, Toronto, New York, and New Orleans. His supervisor has suggested that he use a digital camera to make copies of as much material as he can while doing his fieldwork in the archives, so that hopefully he will not have to do multiple trips to the different cities (his budget for the fieldwork is very limited). He will end up with hundreds, if not thousands, of images of archival documents, programs, posters, and photographs.
- He also plans to interview present and past performers, administrators and Board members of a number of circus companies, and to document a number of performances using a digital video camera. Interviews will be analyzed, possibly using NVivo software, for which the University of Calgary has a site license.
- Lachlan is an aspiring writer and would eventually like to publish a social and pictorial history of circus arts for a general, rather than academic, audience. If he cannot find a publisher prepared to publish this as a book, he might try to get the information out via a website or via his blog, which he also plans to use to promote the project while he is doing it. He has also been approached by the CBC to produce a radio documentary, and plans to use snippets from his interviews as part of this 1-hour show. He thinks the interviews might constitute an interesting oral history collection in their own right and wonders whether Archives Canada or some other institution may be interested in having these at the end of the project.

UNIVERSITY OF CALGARY

Scenario 4

Engineering

- Paul is just starting out on his PhD in Engineering. He is investigating the properties of certain metals in the context of more efficient car design. Paul is interested in pursuing a career as an academic researcher and is more interested in the fundamentals of surface science than he is in cars, but he was pleased to receive a scholarship from the car manufacturer that is supporting the research in the hope that the results will give it a competitive edge.
- Paul is one of four PhD students using this project as the means of completing their PhD - they have the same supervisor, who is the Primary Investigator on the grant that the PhD students are all part of. Paul will be working with samples of various kinds of metals, which will undergo different treatments in the lab. Each student in the lab will be treating the same metals slightly differently and they will need to be able to compare results with each other. The treatment processes vary, and Paul's is one of the most complex - it can take him up to a month to generate a very small number of samples.
- The treated samples will be run through a scientific instrument that produces very large images and lots of them - one experiment might generate hundreds of images. This piece of scientific equipment is provided by a commercial supplier, who also licenses the software needed to perform the analysis and visualisation on the images. The machine has been in use in the department for a while and is pretty slow: there has been talk that it will be upgraded sometime soon, which everyone is really looking forward to as this will speed up the research.
- The second stage of Paul's research will be to model the effects on car efficiency of using metals that have received the treatments. The car manufacturer that is sponsoring his research has a computer model that they have developed themselves and want to validate. Paul will feed his lab-generated data into the models, producing new derived data that may point to design changes that the company could make to improve the efficiency of their vehicles.
- It is likely that prototype cars made from the new materials might be produced as a result of this work, but this would probably not happen in the timeframe that Paul is doing his PhD (he is aiming to complete in 3 years, but the project has at least 5 years of funding). When he finishes, Paul thinks he will seek a post-doc in another institution, and try to further his work using the data that he has derived during his PhD, perhaps applying the findings to another area of transport manufacturing (e.g. high speed rail).

Medicine

Data Collection

- Types of data
  - Interview data (audio recordings, transcripts, NVivo coding)
  - Policy documents (word docs, pdfs)
  - WHO A(H5N1) dataset (pdf) -> derivative data format (csv)
- File formats
  - Audio: mp3
  - Documents: rtf & pdf
  - NVivo: proprietary format, xml exports
- Naming Procedures
  - Interview files named by date (YYYYMMDD), subject number, interviewer, location, version.
  - Documents name by year of publication (date retrieved if not known), year of publication, source institution, author, title.
  - Datasets named by dataset date, acquired data, and WHO dataset title

Medicine

Documentation & Metadata

- What documentation is needed?
  - Description of interview format/procedure, subject information*, parameters & procedures for coding the quantitative data within NVivo.
  - Description of survey technique for gathering documents. List of all documents with notes on documents that cannot be shared.
  - Data dictionary/description file for WHO data (hopefully this exists, if not will create one).
- Procedure for creating documentation:
  - Single researcher will be creating all documentation.
  - Interview format/procedure will be specific as part of Ethics application.
  - Interviewer will record subject info as part of informed consent at time of interview.
  - Will export as much as possible when NVivo analysis is finished as well as any at the time of submission of publication(s).
  - Document list & survey procedure will be documented at the time the survey is put together. This list will be checked for discrepancies at time of publication(s) and end of project.
  - WHO data will be documented at time of analysis.
- Metadata standard
  - Not making use of a specific metadata standard. Documentation documents described above will be saved in rtf format.

Medicine

Storage & Backup

- Storage requirements (space & time)
  - Audio requires approx. 100MB / hr. Estimate 50 hours of interviews so 5GB space. Original audio will be kept until end of project (5 years) and then destroyed.
  - All other data will require 500 MB. Will be kept on lab systems during life of project. Archival system afterwards for at least 10 years.
- How and where storing data:
  - Personal laptop and dropbox during collection
  - Uploading as soon as possible (within a week) into lab's sharepoint environment that is backed up by IT.
- Team & Collaborators
  - Sharepoint provides mechanism to share data with supervisor.

Medicine

Preservation

- Long-term preservation
  - Will be stored with lab's long term data store.
- Is data preservation ready?
  - Aside from NVIVO files, all others are non-proprietary data types.
  - All metadata will be stored with data.

Medicine

Sharing & Reuse

- What data will you be sharing
  - Where allowed by interviewee informed consent, will share interview transcripts and metadata via institutional data repository.
  - No other data will be publicly shared.
- End-user License
  - Data released with CC0 license terms
- How will this be promoted to research community
  - UCalgary's data repository supports OAI harvesting as well as the Canadian Federated Data Repository Discovery tool. This repository also provides a permanent DOI and is indexed by Google.

Medicine

Responsibilities & Resources

- Who is responsible?
  - Louise responsible for managing all data.
- What happens with changes in personnel?
  - If incapacitated, PhD supervisor will have access to everything uploaded to sharepoint.
- What resources require/cost of data management?
  - Given the project costs of data management are relatively minor.
  - No additional expenses necessary to place data in repository.

Medicine

Ethics & Legal Compliance

- Securement management of sensitive data
  - Data is kept on an encrypted laptop as well as in sharepoint where UCalgary username/passwords are required and access is limited to Louis & her supervisor.
- Secondary Uses?
  - Informed consent provides permission to share. Secondary use not a concern.
- Legal, Ethical, IP Issues?
  - Covered by informed consent.
  - WHO data licenses permissions will be checked before use.

Data Management Planning Resources

Lib Guide: Research Data Management @ UCalgary

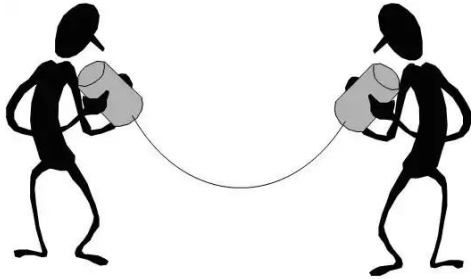
- http://libguides.ucalgary.ca/researchdatamanagement

Other Resources:

- https://libraries.mit.edu/data-management/
- https://library.carleton.ca/sites/default/files/find/data/surveys/pdf_files/Research Data Management - Intro.pdf
- https://www.library.ualberta.ca/research-support/data-management
- http://researchdata.library.ubc.ca/plan/
- https://www.lib.umn.edu/datamanagement
- http://thedata.org/book/data-management-plan-suggested-outline
- https://data.library.virginia.edu/files/DMDocumentation-and-Metadata-for-Eng-and-PhySci_Sept2013.pptx

Questions?

- John Brosz
  - Coordinator, Research Data & Visualization
  - jdlbrosz@ucalgary.ca
- Kathryn Ruddock
  - Manager, Digitization and Repository Services
  - digitize@ucalgary.ca



- University of Calgary's Data Repository
  - https://dataverse.scholarsportal.info/dataverse/calgary
- Portage's DMP Assistant
  - https://assistant.portagenetwork.ca/

Slides: http://dx.doi.org/10.5072/PRISM/31903