Research Data Management

Enhanced Research Skills Certificate Program

Week 3: Understanding Data

Winter 2016

After this workshop you will be able to:

- Be able to explain what data is, and the importance of research data management
- Be able to use a file-naming system to name and organize data files
- Describe the value of open access to data, and what is required to facilitate discovery, access, and reuse of shared data
- Begin to practice good methods of data storage and backup
- Read and interpret a basic metadata record
- Begin recording metadata to accompany the data you collect and use for research

Video:

Data Sharing and Management Snafu in 3 Short Acts

Reflection questions

- What were the issues preventing the researcher from reusing the data?

- Why is it important for researchers to document, preserve and share their data?

But what do we mean by *Data*??

Defining ‘Data’

“Any information that can be **stored in digital form**, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc.

Such data may be generated by various means including **observation**, **computation**, or **experiment**.”

National Science Board. 2005. Long-Lived Data Collections: Enabling Research and Education in the 21st Century. Arlington, VA: National Science Foundation, p. 13.

What is data?

Any information you use in your research

arXiv.org e-Print archive - Mozilla Firefox

File Edit View History Bookmarks Tools Help

arXiv.org e-Print archive

arxiv.org

We gratefully acknowledge support from University of Cambridge

Cornell University Library

Login

(Help | Advanced search)

Search or Article-id

All papers Go!

arXiv.org

Open access to 706,656 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics

Subject search and browse: Physics Search Form Interface Catchup

29 Jul 2011: arXiv Sustainability Initiative Update: An update on the effort to design a new collaborative support model for arXiv, including a short FAQ, is provided in the July 2011 update.

25 Apr 2011: Data sets accepted as part of Data Conservancy pilot; Science

See cumulative "What's New" pages. Read robots beware before attempting

Physics

- Astrophysics (astro-ph new, recent, find)
  includes: Cosmology and Extragalactic Astrophysics; Earth and Pl
  Instrumentation and Methods for Astrophysics; Solar and Stellar A
- Condensed Matter (cond-mat new, recent, find)
  includes: Disordered Systems and Neural Networks; Materials Sc
  Soft Condensed Matter; Statistical Mechanics; Strongly Correlat
- General Relativity and Quantum Cosmology (gr-qc new, recent
- High Energy Physics - Experiment (hep-ex new, recent, find)
- High Energy Physics - Lattice (hep-lat new, recent, find)
- High Energy Physics - Phenomenology (hep-ph new, recent,
- High Energy Physics - Theory (hep-th new, recent, find)
- Mathematical Physics (math-ph new, recent, find)
- Nuclear Experiment (nucl-ex new, recent, find)
- Nuclear Theory (nucl-th new, recent, find)
- Physics (physics new, recent, find)
  elerator Physics; Atmospheric and Oceanic P

http://www.lib.cam.ac.uk/

Defining Sample Records

The principal activity at this stage is to define a sample set of files from the collection which represent the range of characteristics within that whole collection. The recommended sample size is 2-10 files. These files are then uploaded into Plato. Once a file has been uploaded, its file type is identified automatically using JHOVE and FITS. If these tools suggest more than one file type, the user should confirm the choice of file type. It is at this point that XCDL descriptions of the files can be produced, if these are required by the user. XCDL is an Extensible Characterisation Definition Language written as part of the development of Plato (there is an associated Extensible Characterisation Extraction Language, XCEL), and it aims to allow for comparison of files.

Identifying sample records

While ultimately intended to run a sample of the recommended size, for the initial assessment we decided to use only one document, chosen at random. The decision to use only one file was made so that the evaluation process was speeded up, and for us to gain an initial understanding of the programme.

Identify Requirements

The key process here is setting up an appropriate objective tree. This needs to contain all the attributes of the migrated document that need to be compared to the original (e.g. appearance, coherence); other elements of the action path should also be considered at this point.

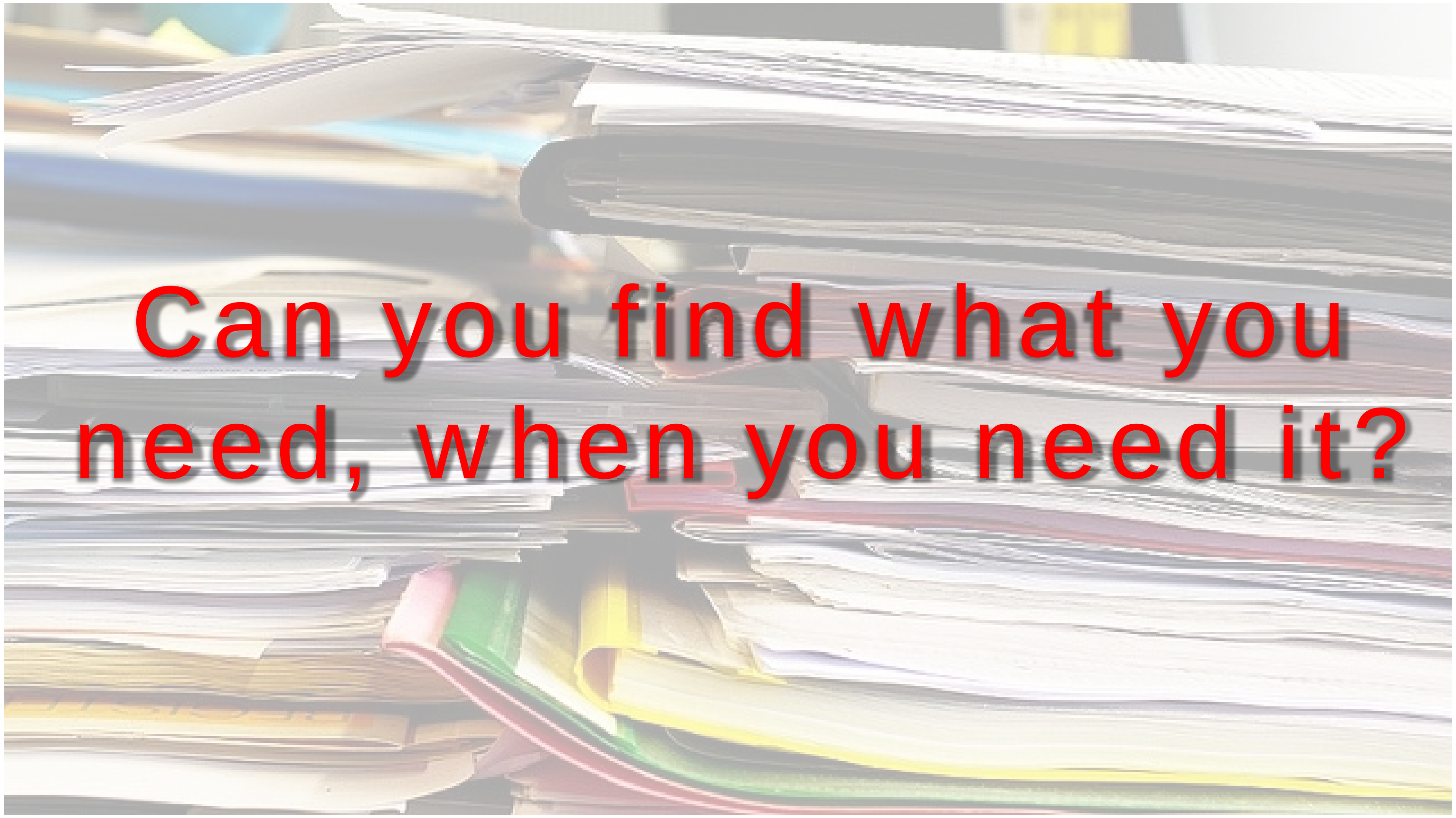
While this issue will be discussed more fully later (see Creating and evaluating the decision trees), it is worth emphasising here the importance of distinguishing between factors that are important and how important they are. Our experience here suggests that, when in doubt, it is better to ascribe Ordinal criteria rather than Boolean criteria for attributes. However, if meaningful numerical criteria can be applied then this may be preferable to Ordinal assessments which may involve a degree of subjectivity. The Transforming stage allows criteria to be applied more or less strictly, depending on context.

Defining the objective tree

The objective tree we used at this stage is given in Appendix B. It is based on the attributes discussed in the interviews with DSpace@Cambridge users, and is therefore indirectly based on the objective tree used by Becker et al. (2007).

We originally used a 0-5 integer range as assessment criteria for most attributes, with 0 indicating that no change was acceptable, and 5 indicating that changes could be made; this was chosen to allow us to align the values we assigned at this stage with the important attributes identified from our interviews with users. However, we soon amended this to an ordinal scheme, to give increased clarity on the range being assessed.

Slide adapted from the
PrePARe Project



Can you find what you need, when you need it?

‘What a mess’ by .pst, via Flickr: http://www.flickr.com/photos/psteichen/3915657914/.

Why Manage Data: Researcher Perspective

- Manage your data for yourself:
  - Keep yourself organized
  - Track your science processes for reproducibility
  - Better control versions of data
  - Quality control your data more efficiently

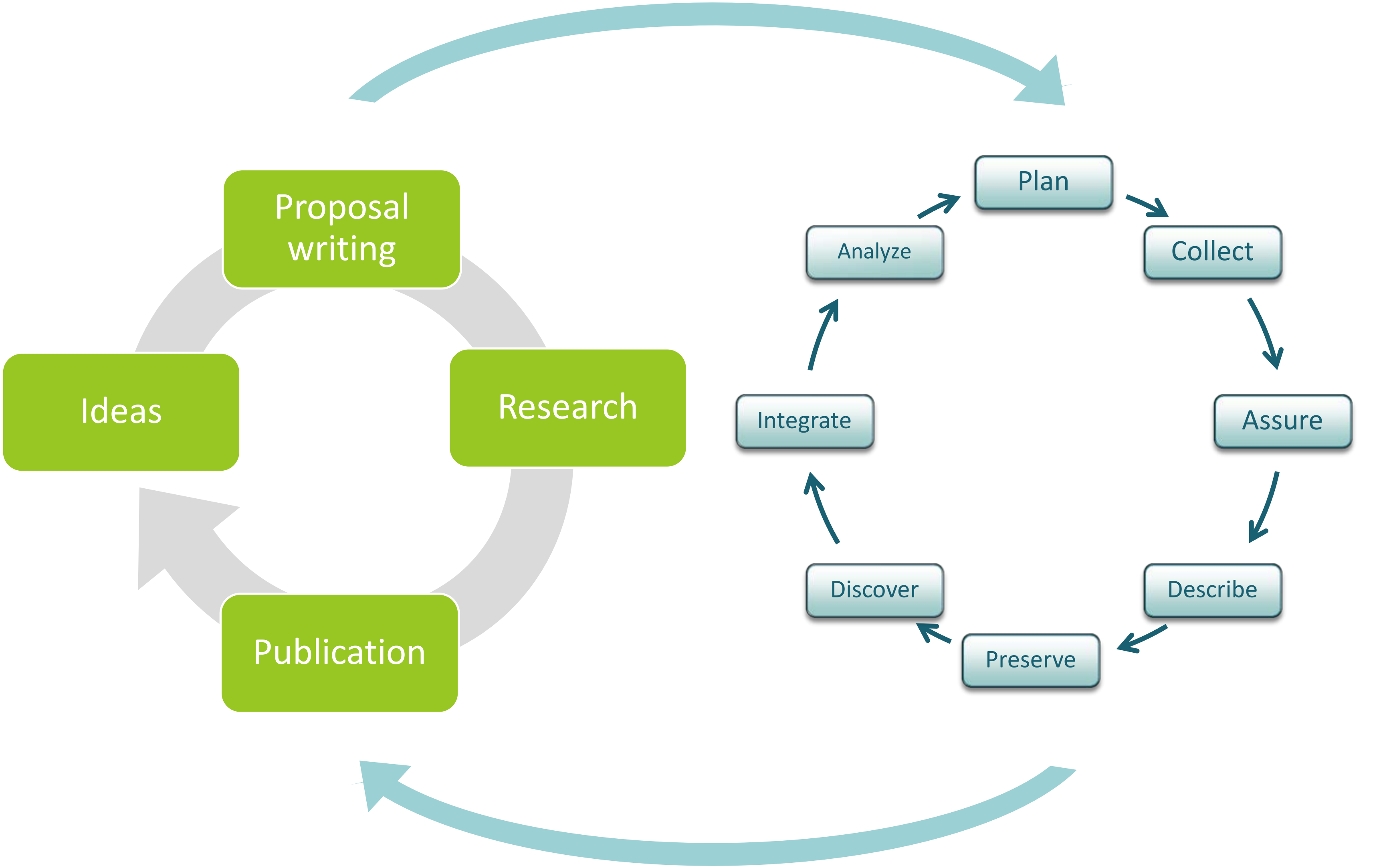
Why Manage Data: Researcher Perspective

- Make backups to **avoid data loss**
- Format your data **for re-use (by yourself or others)**
- Be prepared: Document your data **for your own recollection, accountability, and re-use** (by yourself or others)
- Prepare it **to share it – gain credibility and recognition** for your research efforts!

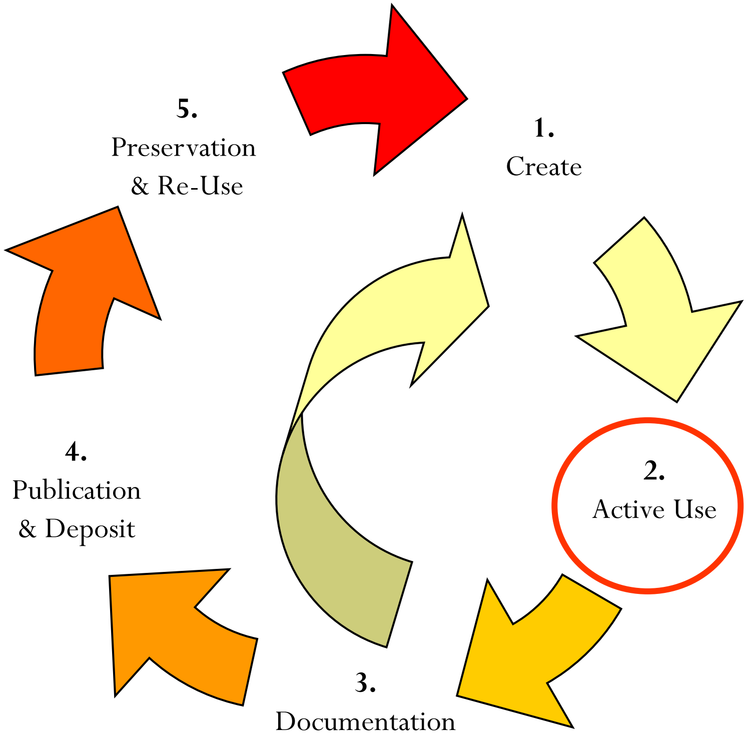
Disciplines and Data Management

- Need for data management is universal
- Practices vary across disciplines
- Good data management extends knowledge creation across disciplines

Research and Data Management Planning are Both Inter-related Iterative Processes

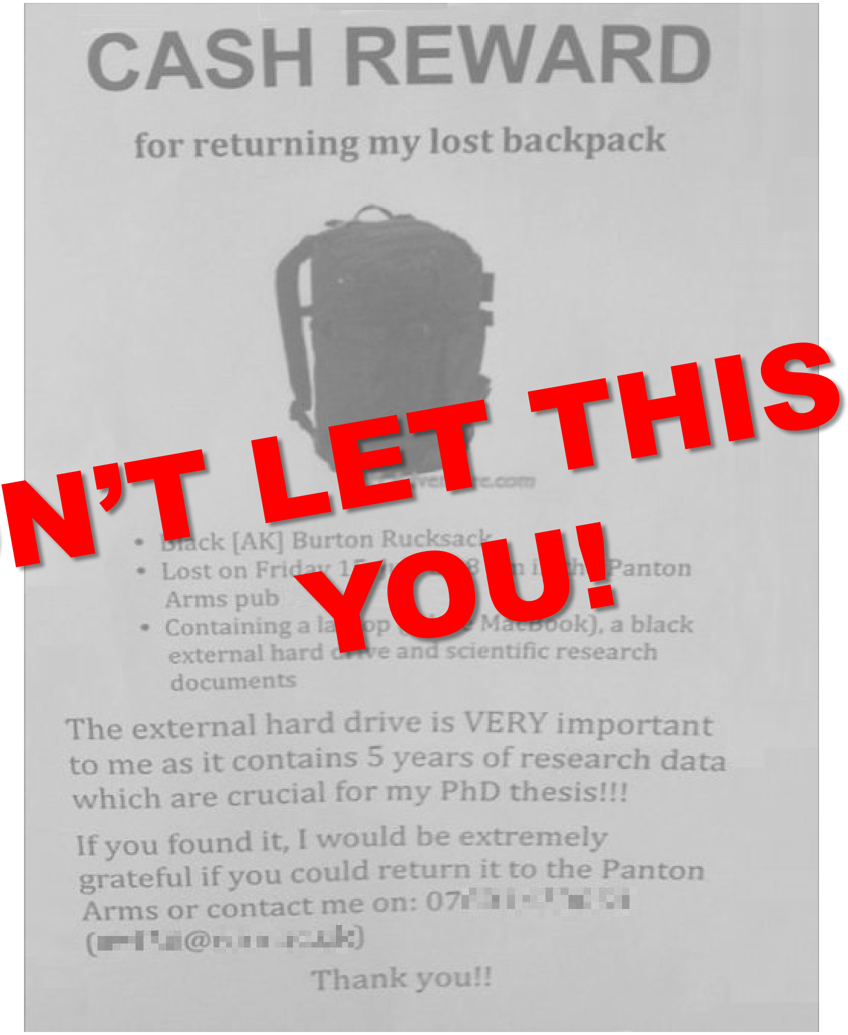


How will you look after your data?



- Is your data safe?
- Is your data organised?
- Can you find your data?

DON’T LET THIS BE YOU!



CASH REWARD

for returning my lost backpack

- Black [AK] Burton Rucksack
- Lost on Friday 15 [illegible] 8 [illegible] in the Panton Arms pub
- Containing a laptop [illegible] MacBook), a black external hard drive and scientific research documents

The external hard drive is VERY important to me as it contains 5 years of research data which are crucial for my PhD thesis!!!

If you found it, I would be extremely grateful if you could return it to the Panton Arms or contact me on: 07[redacted] ([redacted])

Thank you!!

http://blogs.ch.cam.ac.uk/pmr/2011/08/01/why-you-need-a-data-management-plan/

Slide adapted from the
PrePARe Project

Storage and Security

Photo credits: Harvey Rutt
http://www.ecs.soton.ac.uk/regenesis/pictures/

- **3… 2… 1… Backup!**
  - at least 3 copies of a file
  - on at least 2 different media
  - with at least 1 offsite
  - Do not use Dropbox and be very cautious of cloud storage for sensitive data.

- **Access**
  - Protect your hardware
  - For sensitive data, use file encryption (Talk to your advisor)
  - Keep passwords safe (e.g. Keepass)
  - At least 2 people should have access to your data

Data Loss

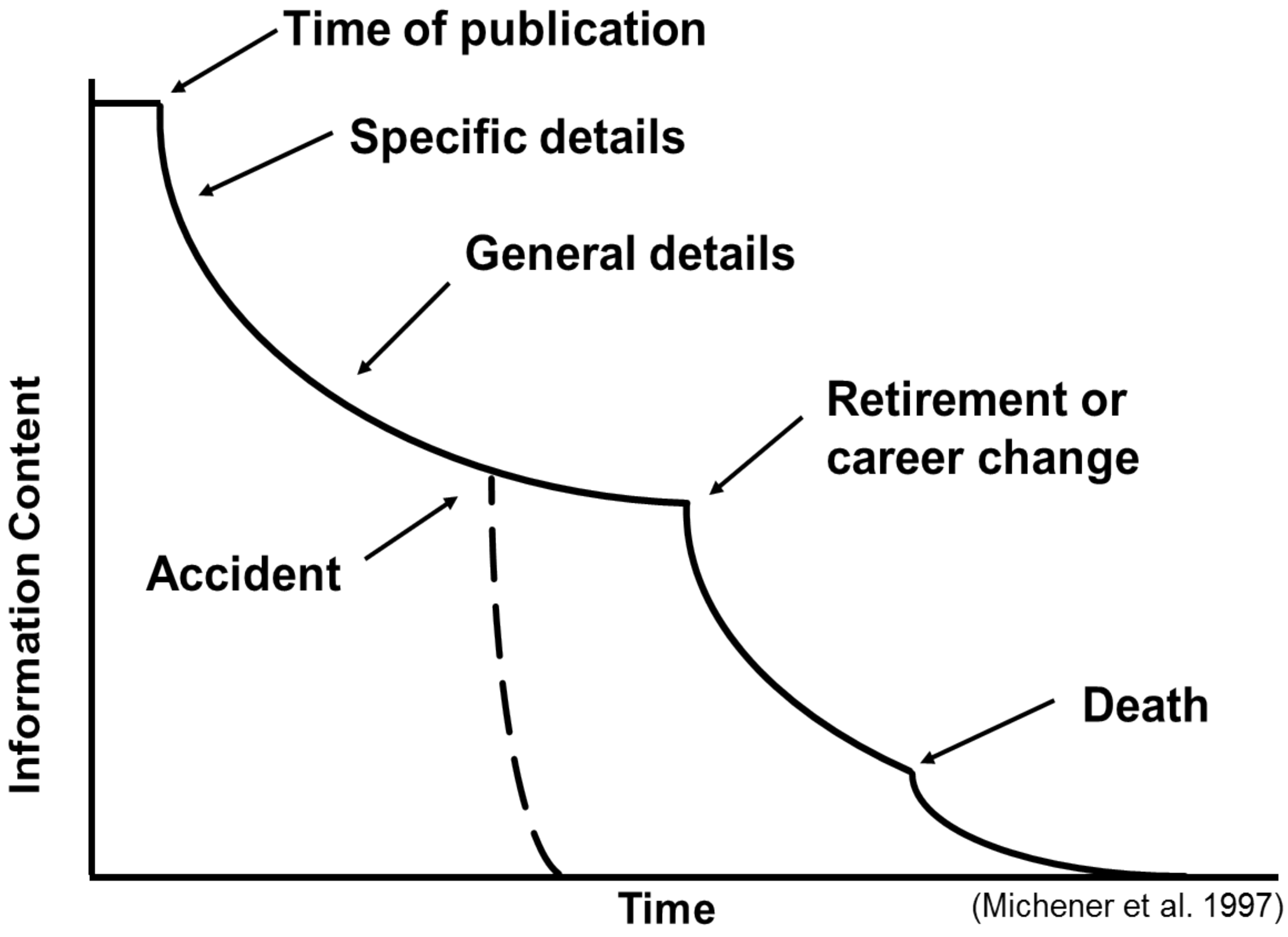
CC image by Sharyn Morrow on Flickr

CC image by momboleum on Flickr

- Natural disaster
- Facilities infrastructure failure
- Storage failure
- Server hardware/software failure
- Application software failure
- Format obsolescence
- Legal encumbrance
- Human error
- Malicious attack by human or automated agents
- Loss of staffing competencies
- Loss of institutional commitment
- Loss of financial stability
- Changes in user expectations and requirements

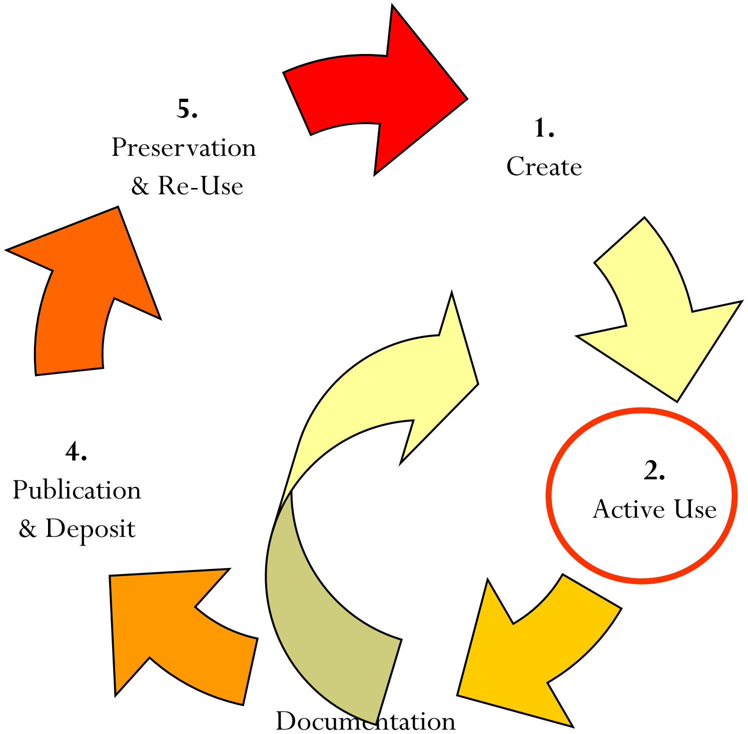
Slide adapted from DataONE Education Module: Why Data Management. DataOne. Retrieved March 21, 2013

Loss does need not be catastrophic, it can be 'entropic'



(Michener et al. 1997)

How will you look after your data?



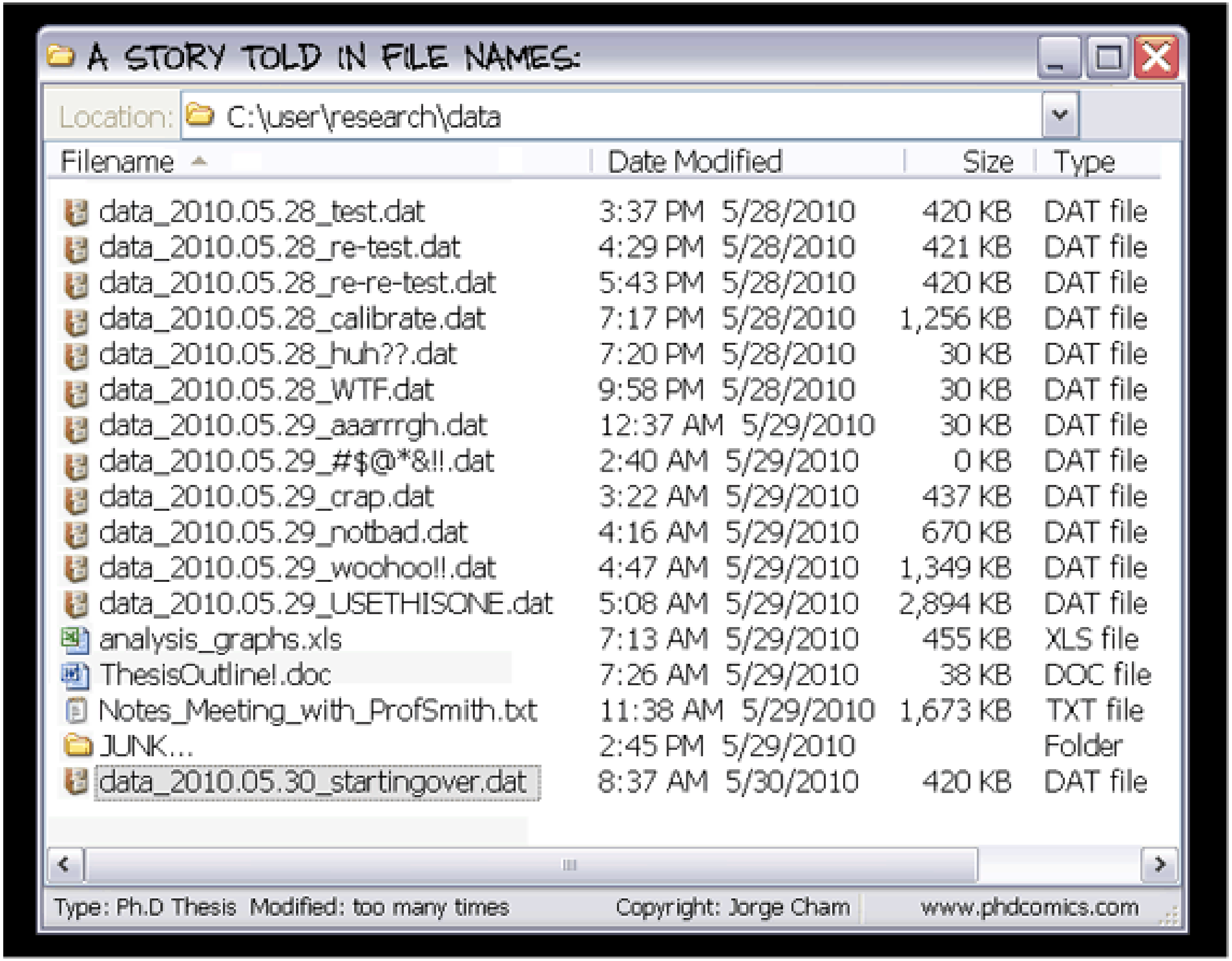
- Is your data safe?
- Is your data organised?
- Can you find your data?

UNIVERSITY OF OREGON

library.uoregon.edu/datamanagement/

The Benefits of Organizing Files and Folders

- Future-proof your data!
- Meet funder obligations
- Save time
- Easier collaboration; ensure that you work on correct version
- Good research practice
- Archive correct version



A STORY TOLD IN FILE NAMES:

Location: C:\user\research\data

Filename	Date Modified	Size	Type
data_2010.05.28_test.dat	3:37 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM 5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM 5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM 5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM 5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrgh.dat	12:37 AM 5/29/2010	30 KB	DAT file
data_2010.05.29_#$@*&!!.dat	2:40 AM 5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM 5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM 5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!!.dat	4:47 AM 5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM 5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM 5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM 5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM 5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM 5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT file

Type: Ph.D Thesis Modified: too many times Copyright: Jorge Cham www.phdcomics.com

File Naming

File Naming Conventions

- Describes essential facts
- Separates files by key data points
- Useful sorting/browsing arrangement

Example:

20120613_southbend_s23.txt

[date]_[location]_[instrument].[extension]

File Naming: Best Practice

- Brief and relevant
- No special characters, dots or spaces
- For separation use underscores _
- Name independent of location
- Date: YYYY_MM_DD

File Naming

File name = **principal identifier** of file

Easy to: identify, locate, retrieve, access

Provides context e.g.:

✓ version number e.g. *FoodInterview_v1*

FoodInterview_v2

✓ date e.g. *HealthTest_2011_04_06*

✓ content description e.g. *BGHSurveyProcedures*

✓ creator name e.g. *CommsPlanHLJ*

File Naming: Have a System!

- **Consistent** and logical naming system
- Develop a system with colleagues for shared data
- Be selective about what you save!
- Prepare it for your future self

File Naming Exercise

Which of the following filenaming conventions would you use, and why?

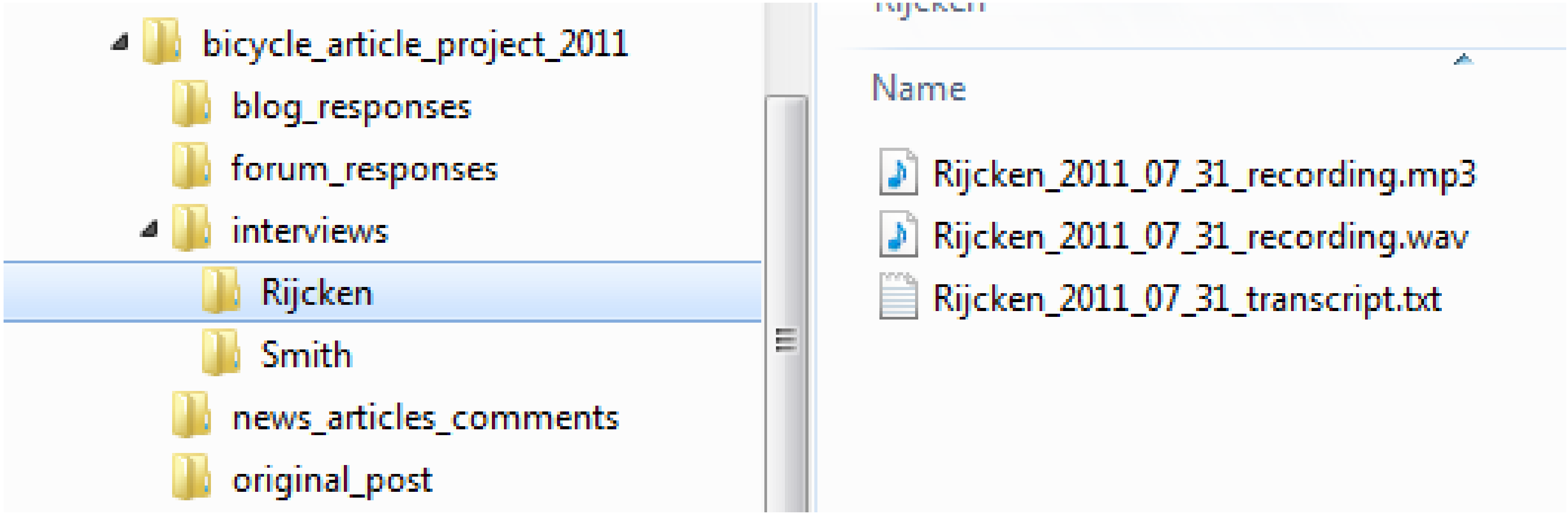
File naming strategies – examples

> In retrospect I am not very happy with the method I used for naming files. The biggest problem was with the newspaper articles I downloaded… I named the files only based on the topic of the article, without mentioning the name of the periodical and the year of publication, which would have been very useful later, when I began writing the thesis.
>
> – Doctoral student researching communication history

Folder Organization

Folder Arrangement Conventions

- Organize project content into discrete units
- Standardize subfolder arrangement



- bicycle_article_project_2011
  - blog_responses
  - forum_responses
  - interviews
    - Rijcken
    - Smith
  - news_articles_comments
  - original_post

Rijcken

Name

- Rijcken_2011_07_31_recording.mp3
- Rijcken_2011_07_31_recording.wav
- Rijcken_2011_07_31_transcript.txt

What Format Should I Store my Data in? (Open Formats)

- What is an Open Format?
  - The specification has been published
  - The format can be used and implemented by anyone
- Advantages of Open Formats
  - Not limited to one piece of software
  - More chance of being able to use the format in the future
- Examples from UO Data Management pages

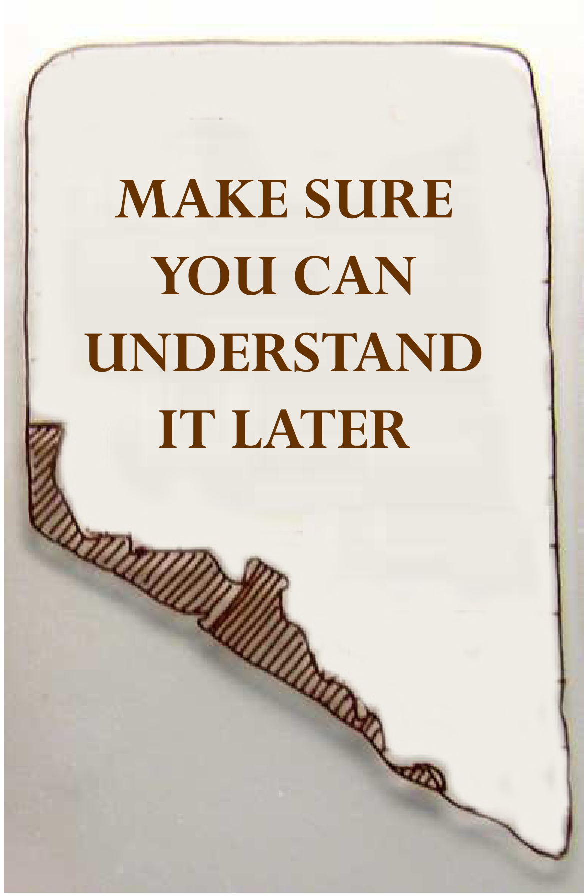
In small groups…

- What data management challenges have you encountered?
- What strategies have you personally found useful?
- Be ready to feed back to the group

- Use the "Nine simple ways…" article for ideas.

Make material understandable

What's obvious now might not be in a few months, years, decades…



Adapted from 'Clay Tablets with Linear B Script' by Dennis, via Flickr: http://www.flickr.com/photos/archer10/5692813531/

Slide adapted from the
PrePARe Project

Make material verifiable

- Detailing your methods helps people understand what you did
- And helps make your work reproducible
- Conclusions can be verified

Image by woodleywonderworks , via Flickr: http://www.flickr.com/photos/wwworks/4588700881/

Understanding your data

- Will you be able to write up your methods at the end of your studies?
  - Will you be able to respond to reviewers' comments?
  - Will you be able to find the information you need for final project reports?
- Can you reproduce your work if you need to?
- What information would someone else need to replicate your work?

Understanding your data

- Do you know how you generated your data?
  - Equipment or software used
  - Experimental protocol
  - Other things included in (e.g.) a lab notebook
  - Can reference a published article, if it covers everything
- Are you able to give credit to external sources of data?
  - Include details of where the data are held, identified & accessed
  - Cite a publication describing the data
  - Cite the data itself e.g.

Gary King; Langche Zeng, 2006, "Replication Data Set for 'When Can History be Our Guide? The Pitfalls of Counterfactual Inference'" hdl:1902.1/DXRXCFAWPK UNF:3:DaYlT6QSX9r0D50ye+tXpA== Murray Research Archive [distributor]

What is Metadata?

Metadata is: Data 'reporting'

- **WHO** created the data?
- **WHAT** is the content of the data?
- **WHEN** were the data created?
- **WHERE** is it geographically?
- **HOW** were the data developed?
- **WHY** were the data developed?

Photo by Michelle Chang. All Rights Reserved

Metadata Exercise

Look at one of these data records with your group:

Ecology: http://bit.ly/1ToAb9d

Psychology: http://openpsychologydata.metajnl.com/articles/10.5334/jopd.e2/

Please answer the following questions:

a. Is there a license or data use requirement for you to use this dataset?

b. Are the data collection and processing methods described?

c. How would you cite this dataset?

Backups vs. Preservation

What's the difference?

- Backup = Workflow Recovery
- Preservation = Access (long term)

- **Backups**
- periodic snapshots in case current work is destroyed or lost

Backups vs. Preservation

- What's the difference?

- **Backups**
- periodic snapshots in case current work is destroyed or lost

- **Preservation**
- long-term retention for future discovery and use

SKYNET
NEURAL NET-BASED ARTIFICIAL INTELLIGENCE
CYBERDYNE SYSTEMS CORPORATION

Why Share Data?

- Share with your future self – avoid repeating research!
- To fulfill funder requirements. For instance, NIH has outlined key elements to consider in preparing a data sharing plan.
- Some journals and societies require data archiving, i.e., the Dryad partner journals have a Joint Data Archiving Policy (JDAP).
- To raise interest in your research. Sharing detailed research data is associated with increased citation rate.
- To speed research, particularly in complex research fields
- To establish priority and a public record.

How and where do you share* data?

- Data repositories – more next week!

The Dataverse Network™ Project

DRYAD

ICPSR

DataONE
Data Observation Network for Earth

Henry A. Murray Research Archive
at Harvard University

*** and find data, too.**

Take-aways

- Manage data well throughout the lifecycle
- Good practices at the start of research make it easier to follow through at the end of research
- Data sharing improves science, increases citation of your work, and is emphasized by research funders

What steps will you take?

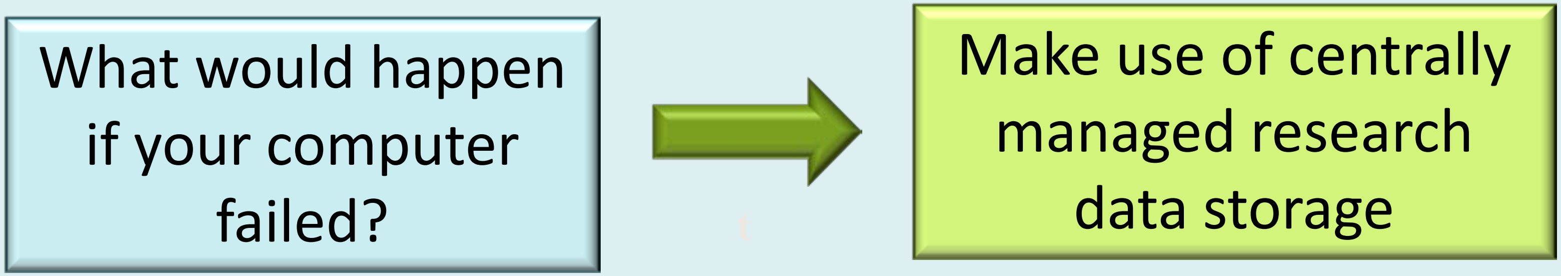
Share with your future self!

Will you be able to write your thesis/ final report in 3 years time?

Use better file names and capture contextual information

What steps will you take?

Share with your future self!

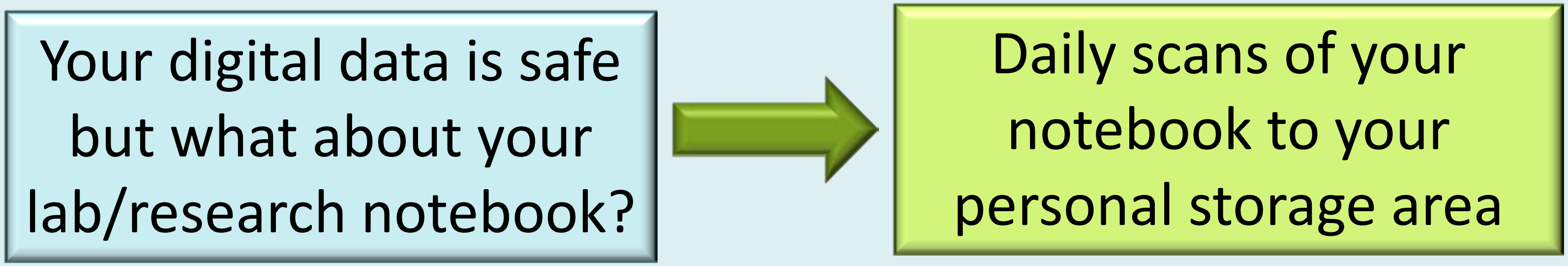


What would happen if your computer failed?

Make use of centrally managed research data storage

What steps will you take?

Share with your future self!

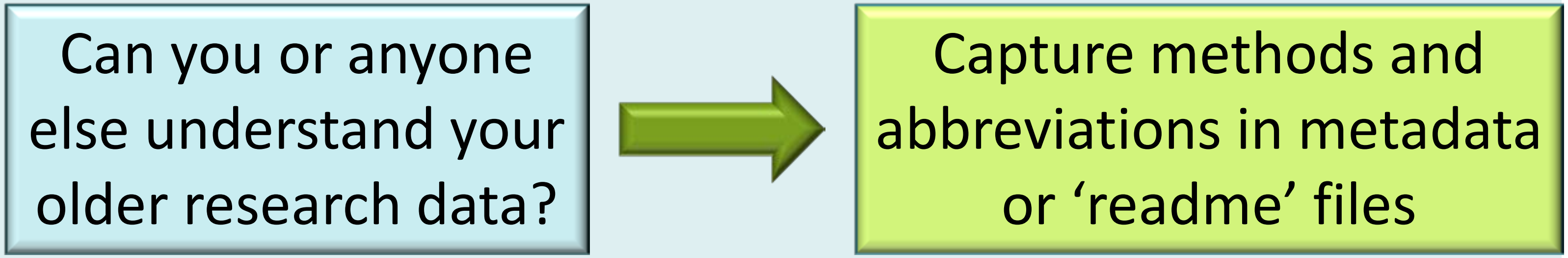


Your digital data is safe but what about your lab/research notebook?

Daily scans of your notebook to your personal storage area

What steps will you take?

Share with your future self!



Can you or anyone else understand your older research data?

Capture methods and abbreviations in metadata or ‘readme’ files

What steps will you take?

Share with your future self!

Does your funder require you to make your data accessible?

Publish your data in a data repository or data journal

References and Sources

- UO Data Management Web site: https://library.uoregon.edu/datamanagement/
- King, Gary. 2011. “Ensuring the Data-Rich Future of the Social Sciences.” *Science (New York, N.Y.)* 331 (6018) (February 11): 719–21. http://apps.webofknowledge.com/full_record.do?product=WOS&search_mode=GeneralSearch&qid=9&SID=3DMGm2g4b4nJISv4h6x&page=1&doc=1.
- Piwowar, Heather A, Roger S Day, and Douglas B Fridsma. 2007. “Sharing Detailed Research Data Is Associated with Increased Citation Rate.” *PloS One* 2 (3): e308. http://www.plosone.org/article/fetchObject.action?uri=info:doi/10.1371/journal.pone.0000308&representation=PDF.

Some slides and materials in this presentation are from:

- Managing Your Research Data, Catherine Pink and Jez Cope, University of Bath, UK
- DaMaRO Project Introduction to Research Data Management - PowerPoint presentation with presenter's notes - June 2013

Questions?