

Research Integrity and Trustworthy Science: Challenges & Solutions

Coffee & Conversation with Ivan Oransky

Graduate students, post-docs, and research trainees are invited to meet informally with Dr. Oransky to discuss research rigor, retractions, and other insights into the scientific process.

- ✓ Co-founder of Retraction Watch in 2010
- ✓ Distinguished Writer in Residence at the NYU Arthur Carter Journalism Institute
- ✓ Managing editor online for Scientific American
- ✓ Deputy editor The Scientist



Thursday, April 19, 2018

12:30 pm – 2:00 pm

Ben Pomeroy Student-Alumni Learning Center, Room 215

College of Veterinary Medicine, St. Paul Campus

Sponsored by : College of Veterinary
Medicine
& College of Biological Sciences



Managing Research Data to Improve Research Reproducibility



Quality Central 

Sharpening the focus on sound science and quality practices

 **LIBRARIES**
UNIVERSITY OF MINNESOTA

MANAGING YOUR DATA

Home

Our Services

1. Before Your Research

2. During Your Research

3. After Your Research Ends

Training and Workshops

About Us

[www.lib.umn.edu/
datamanagement](http://www.lib.umn.edu/datamanagement)

MANAGING YOUR DATA

Got data? We're here to help you manage, share, and preserve your research data. In addition to our [Data Repository for the U of M](#) curation services, the Libraries will help you navigate available campus resources throughout the data lifecycle:



Before Your Research Begins

- Schedule a [data management plan \(DMP\)](#) consultation ([Request Form](#)) or use our [Explore funding agency requirements](#) for data and learn best practices for getting [IRB approval](#) for sharing data.
- See [more tools for planning for data management](#)



During Your Research

- Attend workshops and explore online [training resources on best practices for data management](#)
- [Get help](#) creating documentation and using metadata standards
- Discover appropriate [U of M services for data, such as data storage](#)
- See [more tools for managing your data during your research](#)



After Your Research Ends

- Share your data broadly in the [Data Repository for U of M](#)
- Self-archive [your data in a disciplinary repository](#)
- See [more tools for archiving and sharing your data after your research ends](#)

What counts as data?

Data are...

- ⊙ Lab results
- ⊙ Computer simulations
- ⊙ Survey results
- ⊙ Observations
- ⊙ Interviews
- ⊙ Textual analysis
- ⊙ Audio, video, multimedia
- ⊙ Metadata
- ⊙ And more....

What data aren't...

- preliminary analyses
- drafts of scientific papers
- plans for future research
- peer reviews, or communications with colleagues
- physical objects (e.g., laboratory samples)
- trade secrets or commercial information
- materials necessary to be held confidential by a researcher until they are published, or similar information which is protected under law

Circular No. A-110 - Uniform Administrative Requirements for Grants and Agreements With Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations. (1999, September 30).


What's data management?

A decorative network diagram in the top right corner, featuring a series of interconnected nodes and lines, resembling a molecular structure or a data network.

Managing data involves:

- ◎ Knowing what types of data you have
- ◎ Having a plan for documentation
- ◎ Preventing data loss or unwanted access
- ◎ Sharing (or not sharing) your data as appropriate
- ◎ Keeping for your data accessible in the future

A data management plan (or DMP) is a document that articulates how these goals will be accomplished.

A decorative network diagram in the bottom left corner, featuring a series of interconnected nodes and lines, resembling a molecular structure or a data network.

Why manage data?

Because

- ◎ Data management is tied to the responsible conduct of research
- ◎ Data are very valuable and collecting/generating data is an expensive and time-consuming process
- ◎ Managing data means that data can be reused in new and interesting ways, both by yourself and others
- ◎ You may be legally required to

Federal Agencies with plans to date

- Department of Agriculture (USDA)
- Department of Transportation (DOT)
- Department of Commerce
 - National Institute of Standards and Technology (NIST)
 - National Oceanic and Atmospheric Administration (NOAA)
- Department of Defense (DOD)
- Department of Education
 - Institute of Education Sciences (IES)
- Department of Energy (DOE)
- Environmental Protection Agency (EPA)
- National Aeronautics and Space Administration (NASA)
- National Science Foundation (NSF)
- Department of Health & Human Services (HHS)
 - Food and Drug Administration (FDA)
 - Center for Disease Control and Prevention (CDC)
 - Administration for Community Living (ACL)
 - Agency for Healthcare Research & Quality (AHRQ)
 - Assistant Secretary for Preparedness and Response (ASPR)
 - National Institutes of Health (NIH)
- Department of Veterans Affairs (VA)
- Smithsonian Institution
- US Geological Survey (USGS)

It's not just federal funders...

© Private funders

- Bill and Melinda Gates Foundation
- Sloan Foundation
- Ford Foundation

© Journals

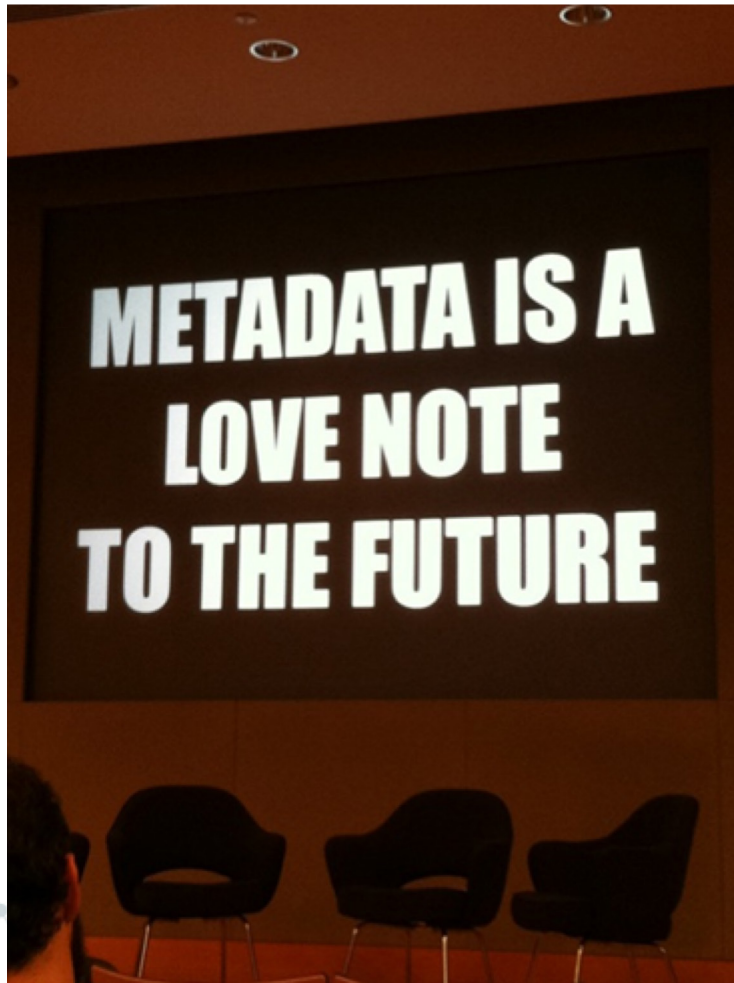
- *Science*
- *Nature*
- *PLoS ONE*
- *PNAS*

And many others...

The background of the image is a light gray network graph. It consists of numerous small circular nodes, some of which are solid gray and others are hollow with a gray outline. These nodes are interconnected by a web of thin, light gray lines representing edges. The overall pattern is dense and non-uniform, filling the entire frame.

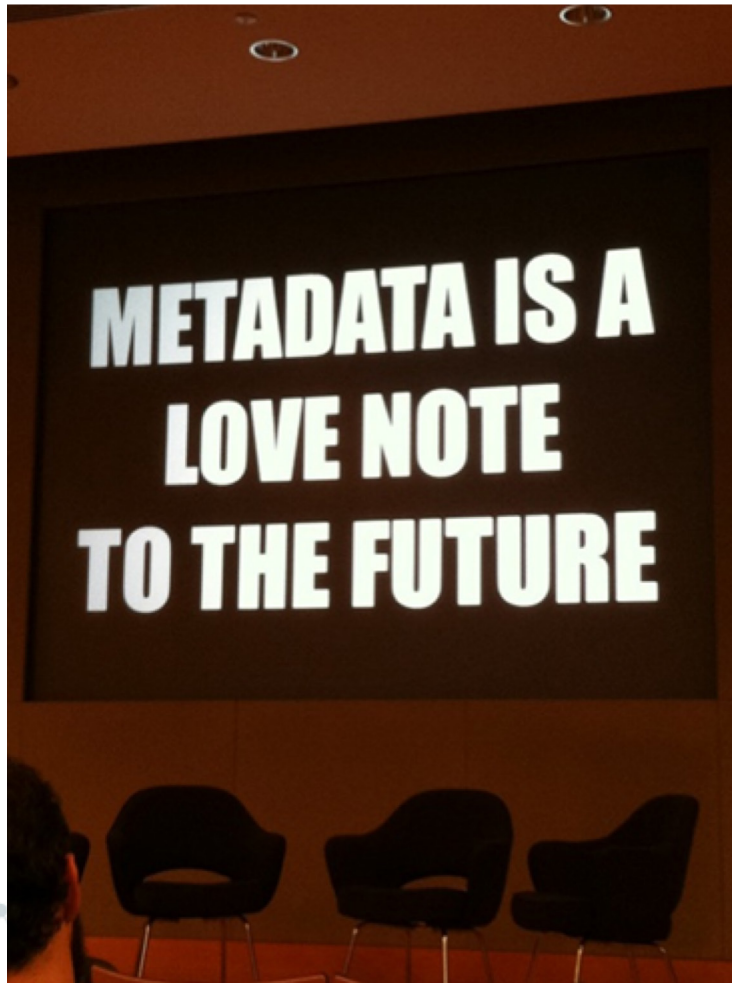
Fatalism is useful

Because human beings forget



- ◎ Data inventories
- ◎ Strategies for documentation
- ◎ File naming conventions
- ◎ Folder structure and organization

Metadata: Data about Data



- ◎ Describes and provides context for data
- ◎ Metadata schemes = project level metadata
- ◎ File properties (title, creator, date) = file level metadata
- ◎ Spreadsheet headings, labels = item level metadata

Good
Documentation
Practices (GDPs):
a systematic
approach to
preparing,
approving, storing,
and archiving
documentation



	University of Minnesota	Doc. No.: QCen.REF.032
	Quality Central	Revision No.: 1
		Approval Date: 5/18/2017
		Page 1 of 3
Title: Research Quality Assurance (RQA) Requirements		

Scientist Name/Role	Department	PI	Project	Date

1	PROJECT MANAGEMENT	Examples of Documentation
	A Project Plan exists	
	Procedures are in place for periodic audit/assessment to monitor compliance with RQA best practices.	
2	PERSONNEL	Examples of Documentation
	A current organization chart illustrates the research roles and reporting lines.	
	Personnel are qualified (education, experience) to perform tasks and initial/ongoing training is conducted and documented.	
	Job descriptions, signature/initial logs, resumes and CVs are current for research staff and students.	
3	FACILITIES	Examples of Documentation
	Procedures exist to keep the lab and storage areas clean and orderly to reduce risk of contamination.	
	Security procedures are implemented to prevent the access of unauthorized personnel.	
	Storage areas are provided for the safe and secure storage of all documentation and also for samples and materials associated with research.	
	A Lab Safety plan exists and safety audit reports are maintained.	
	Environmental control systems are available if needed to provide a stable environment for the research being conducted.	
4	EQUIPMENT	Examples of Documentation
	An equipment inventory list is maintained.	
	Financial support (budgeted) for equipment care has been established.	
	Equipment use records are maintained.	
	Equipment manuals are available.	
	SOPs or manuals address methods, materials, schedules for:	
	a. Routine inspection (e.g., safety, verification)	
	b. Preventive Maintenance	
	c. Operation (e.g., set-up, clean-up, use)	
	e. Calibration and/or standardization	

Basics of Documentation

◎ At a minimum, note:

- How and where you got that piece of data
- Any changes that were made to clean or refine it
- All analyses conducted and all findings, whether expected or not

◎ Don't just record what was done, record how it was done, why, when, and who did it

- “If it isn't documented, it didn't happen”




Attributable: signed, dated, any changes attributed

Legible: can be easily read and understood

Contemporaneous: data recorded immediately as generated

Original: original data, or if not original, location of the original source is included and accuracy confirmed

Accurate: clean, objective recording including all contextual and explanatory information



This codebook.txt file was generated on <YYYYMMDD> by <Name>

GENERAL INFORMATION

1. Title of Dataset

2. Author Information

Principal Investigator Contact Information

Name:

Institution:

Address:

Email:

Associate or Co-investigator Contact Information

Name:

Institution:

Address:

Email:

Associate or Co-investigator Contact Information

Name:

Institution:

Address:

Email:

3. Date of data collection (single date, range, approximate date) <suggested format YYYYMMDD>

4. Geographic location of data collection (where was data collected?):

5. Information about funding sources that supported the collection of the data:

Data Inventories

What is it?

- Title
- Author
- Description
- Date created

Who Owns it?

- Department Owner
- Author
- Contact (Responsibility)

Who can access it?

- Restrictions in place
- Users who have access
- Protocol for handling requests

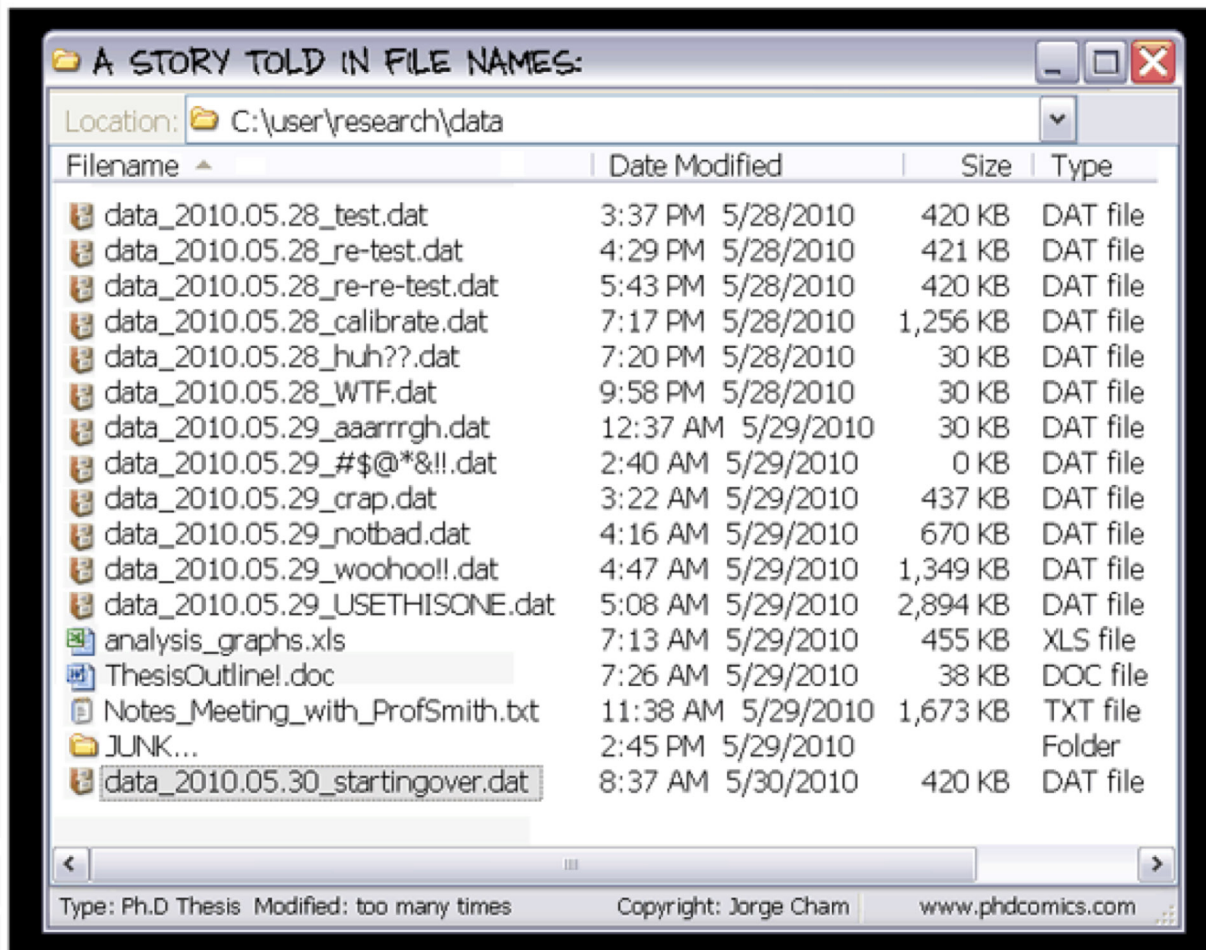
What does it consist of?

- How many files
- Where are they stored
- File size
- Growth Rate
- Are there multiple versions

How Secure is it?

- Uniqueness
- Protections in place
- Back-up locations (2 or more)
- Procedure for format migration
- Retention period


File Naming Conventions



Descriptive File Naming

A decorative network diagram in the top right corner, featuring a series of interconnected nodes (circles) and lines, some solid and some dashed, creating a web-like structure.

Elements to consider when naming files

- ◎ Date
 - ◎ Project name
 - ◎ Type of data
 - ◎ Location/site/spatial coordinates
 - ◎ Researcher info
 - ◎ Version
- 
- A decorative network diagram in the bottom left corner, featuring a series of interconnected nodes (circles) and lines, some solid and some dashed, creating a web-like structure.

Best Practices

◎ Don't rely on nesting in folders
2016/fall/project/data/script.R

◎ Be descriptive
Bad: outline.docx
Better: pubh6646_outline.docx
Best: seminar_pubh6646_notes.docx

Best Practices

- ◎ Use numerical dates
YYYYMMDD rather than Aug15
- ◎ Avoid file names with UPPERCASE letters,
weird characters (/,#?), or spaces between
words

Best Practices

◎ Use consistent structure that falls into useful order (for sorting) and decide if shared terminology is necessary

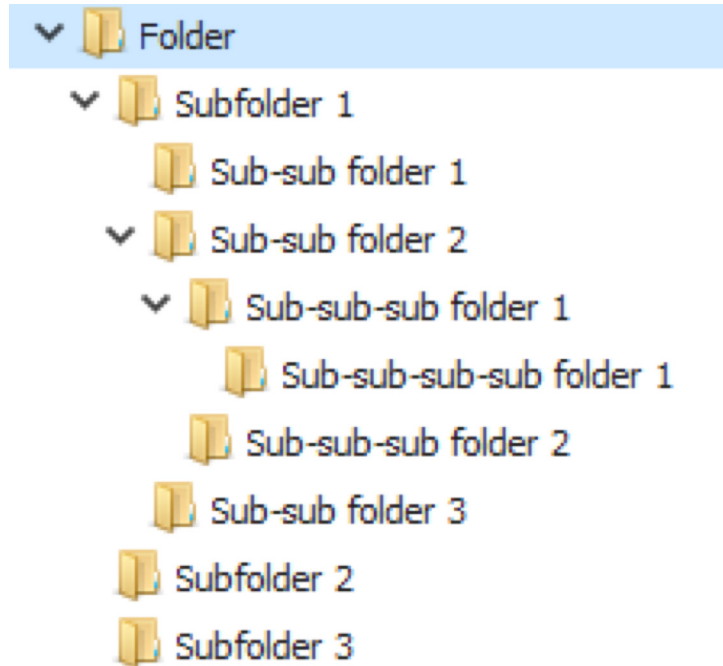
◎ List versions alphanumerically

Bad: draft, finaldraft, reviseddraft

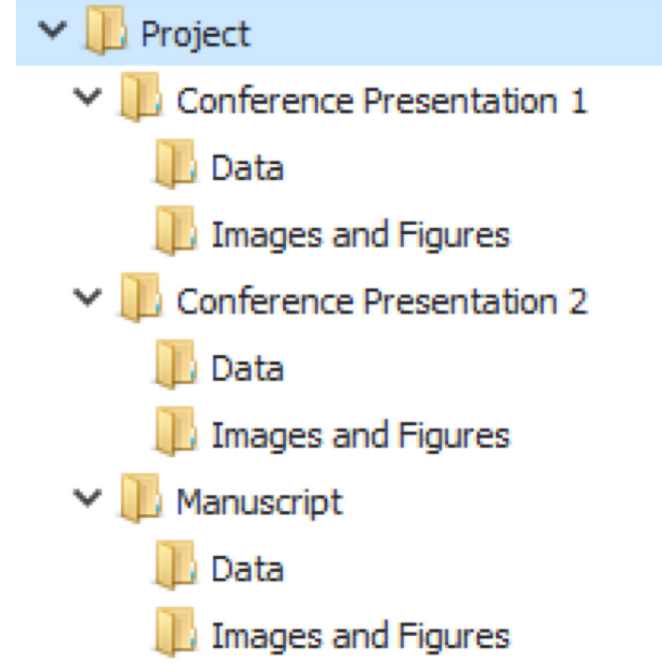
Better: v1, v2, v3

Best: v01, v02, v03

Folder Structures



Be shallow



Avoid duplication

File/Folder Organization

Possible strategies:

- ◎ **By stage:** collection materials, raw data, processed data, shared data
- ◎ **By data type:** databases, text, images, models, etc.
- ◎ **By research activities:** interviews, surveys, experiments, etc.
- ◎ **By materials:** data, documentation, publications, etc.

Because nothing lasts forever



Backing up

Obsolescence

Preservation



3-2-1 Rule

Have 3 copies of your work

On 2 different types of media

With at least 1 remote copy



Storage Options

	Sensitive data	Non-public data	Self-managed collaboration
AHC managed servers			x
Box Secure Storage			
Encrypted Drive/Container			x
Shared/Personal drive	x		x
UMN Google Drive	x		
Dropbox	x	x	
Amazon (personal)	x	x	
Email	x	x	x

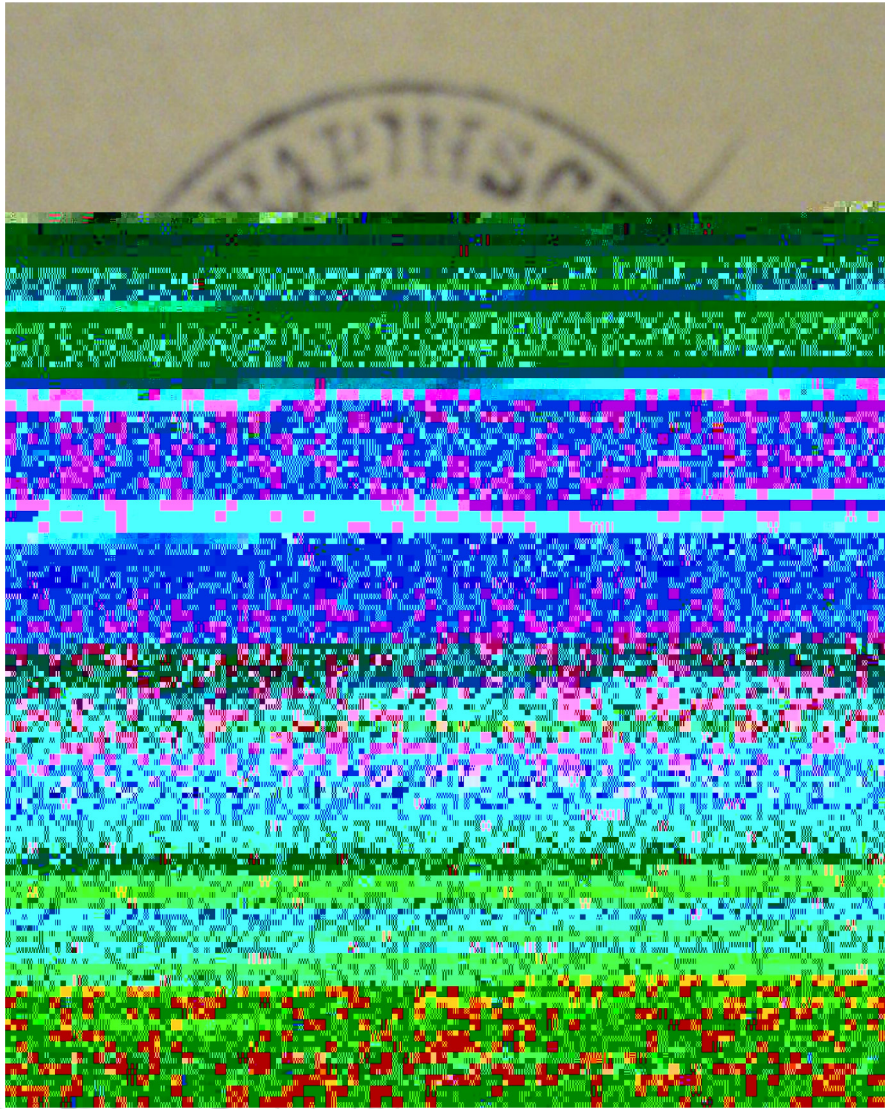
You can back up your data, and still end up with this...





Media Failure: A storage medium no longer functions or functions inadequately

Software Obsolescence: When software no longer runs on current operating systems or is replaced by later versions of software



Bit-level corruption

The data becomes
“scrambled” and
systems can no
longer read it

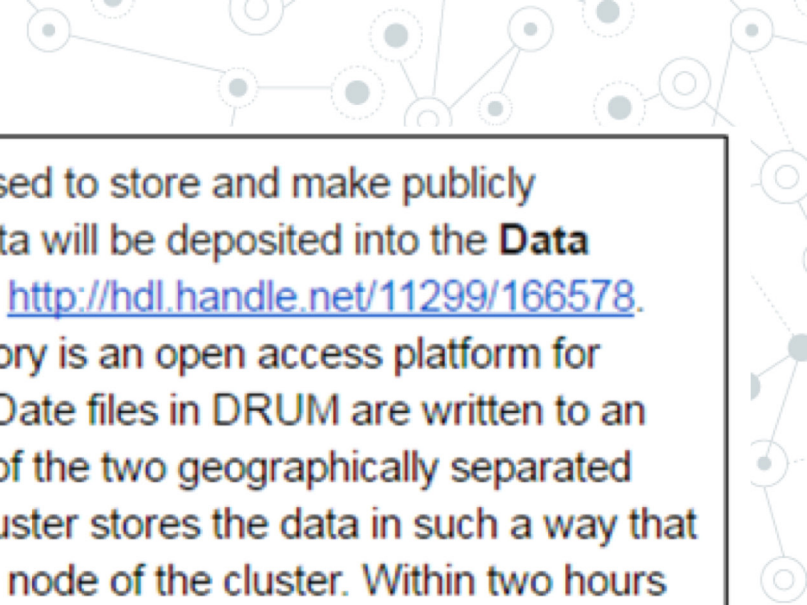
Preservation Steps

- ◎ Only keep what you need
- ◎ Copy your local files onto new storage media every few years
- ◎ Convert your files into open or more accessible file formats
- ◎ Allocate resources to do this preservation work

Digital Preservation

Best practice: an uncompressed, non-proprietary version of data files

For this...	Use this...	Not this...
Text	.pdf, .txt, .html	.docx
Images	.tiff	.psb
Tabular Data	.csv	.xlsx
Video/Media	.mp4	.wmv, .arf, .mov




A long-term data sharing and preservation plan will be used to store and make publicly accessible the data beyond the life of the project. The data will be deposited into the **Data Repository for the University of Minnesota (DRUM)**, <http://hdl.handle.net/11299/166578>. This University Libraries' hosted institutional data repository is an open access platform for dissemination and archiving of university research data. Data files in DRUM are written to an Isilon storage system with two copies, one local to each of the two geographically separated University of Minnesota Data Centers. The local Isilon cluster stores the data in such a way that the data can survive the loss of any two disks or any one node of the cluster. Within two hours of the initial write, data replication to the 2nd Isilon cluster commences. The 2nd cluster employs the same protections as the local cluster, and both verify with a checksum procedure that data has not altered on write. In addition, DRUM provides long-term preservation of digital data files for at least 10 years using services such as migration (limited format types), secure backup, bit-level checksums, and maintains a persistent DOIs for data sets, facilitating data citations. In accordance to DRUM policies, the (deidentified, if applicable) data will be accompanied by the appropriate documentation, metadata, and code to facilitate reuse and provide the potential for interoperability with similar data sets.

Boilerplate language taken from the
Libraries' DMP template



The background of the slide is a light gray network graph. It consists of numerous nodes, represented by small circles, some of which are solid gray and others are hollow with a gray outline. These nodes are interconnected by a web of thin, light gray lines representing edges. The overall pattern is dense and non-uniform, filling the entire background.

**Now (some of)
the trickier stuff...**



HIPAA (Health Insurance Portability and Accountability Act): ensures the privacy of patients' medical information

FERPA (Family Educational Rights and Privacy Act): ensures the privacy of students' educational records

FISMA (Federal Information Security Management Act): ensures the protection of government information and information systems



Privacy & Sensitive Data

There are 18+ HIPAA Identifiers


- 1.Name
- 2.All geographic divisions smaller than a state
3.
 - ⦿ All elements of dates, except the year (includes date of birth, admission date, discharge date, date of death, etc.)
 - ⦿ All ages over 89 (aggregate category of individuals 90 and older)
- 4.Phone number
- 5.Fax number
- 6.E-mail address
- 7.Social security number
- 8.Medical record number
- 9.Health plan number
- 10.Account numbers
- 11.Certificate or license numbers
- 12.Vehicle identification/serial numbers, including license plate numbers
- 13.Device identification/serial numbers
- 14.Universal resource locators (URLs)
- 15.Internet protocol (IP) addresses
- 16.Biometric identifiers
- 17.Full face photographs and comparable images
- 18.Any other unique identifying number, characteristic, or code

Privacy & Sensitive Data

HIPAA variables are direct identifiers (pieces of information that clearly point to one individual)


There are also indirect identifiers

- ◎ Combinations of variables that together can identify an individual
- ◎ No set list of variables, depends on your data set and what other data sets are available



“Data can be either perfectly
useful or perfectly anonymous
but never both.”

Ohm P. Broken promises of privacy: Responding to the surprising
failure of anonymization. *UCLA Law Review*. 2010;57(6):1701-77.



Prior or During Collection

Get consent to share and avoid overly restrictive language in the informed consent process

- Some suggestions for language are available here:
<http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/conf-language.html>

Don't "identify" in the data creation process

- When recording audio, do not use names, place of employment, or other information
- Arrange with participants to use pseudonyms

For audio and video, bleep out names and blur faces or other identifying characteristics

After Data Collection

Some Potential Strategies:

- Eliminate the variable entirely from the dataset

- Recoding variables into broader categories

- Top-coding (restricting the upper range of a variable)

- Match unique cases on the indirect identifier, then exchange the values of key variables between the cases

Best Practice:

- Be cautious of small subgroups and cases with outliers

- Clearly mark any replacements in the data using brackets, tags, or another consistent method

- Keep a secure copy of the non-anonymized data

- Create a log of all the replacements, aggregations, or removals made in each data file. Store this log file separately from the de-identified data

Some Strategies for Securing Data

Authentication (password protected)

- For files, programs, machines, etc.

Firewalls

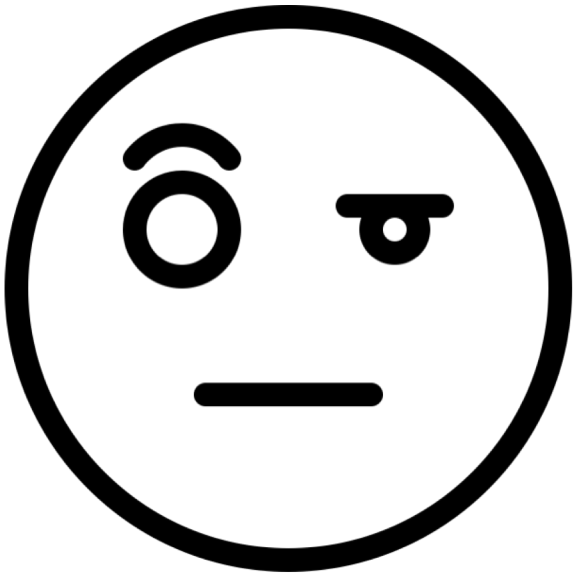
Encrypted files

Physical security

Secure data deletion and secure disposal



Do we really need to know about all this?



- ◎ This (and more) is the type of information covered in a DMP
- ◎ Different elements will be more or less important depending on the type of data
- ◎ Knowing about \neq knowing everything or memorizing

Resources

- © DMPTool: <https://dmp.cdlib.org/>
- © DMP Template & Checklist:
z.umn.edu/dmptemplate
- © readme.txt (project-level documentation):
z.umn.edu/readme
- © OIT Storage & Data Protection Services:
<https://z.umn.edu/19up>



Caitlin Bakker

Research Services Librarian

cjbakker@umn.edu

612-301-1353

Rebecca Davies, PhD

Director, Quality Central

rdavies@umn.edu

612-626-0168





Thank You!

Questions? Comments?

