# Research Data Management

Matthew Murray

Data Librarian

matthew.murray-2@colorado.edu



#### What we'll talk about

- What is research data?
- Data Management vs. Data Management Plans
- Data Storage & Access
- Backups & Versioning
- Documentation
- File Names & Structures
- File Formats & Units of Measurement
- Data Management Plans
- Tools



# No two people on any campus will agree on what "Data" actually is.



- Primary / Secondary
- Qualitative / Quantitative
- Experimental / Observational

- 3D models and printable files
- Accreditation reports
- Archival university papers
- Artistry and performance materials
- Audio
- Books
- Computer code & scripts
- Conference proceedings
- Course catalogs
- Datasets
- Designs & blueprints
- Digital journals
- Dissertations
- Documentation
- GIS files
- Grant proposals
- "Grey" literature
- Historical documents
- Images
- Interviews
- Journals
- Lab notebooks

- Learning materials
- Lecture transcriptions
- Maps
- Methodologies & Workflows
- MOOCs
- Newsletters
- Oral History
- Physical artifacts and specimens
- Point clouds
- Posters
- Presentations
- Seismic recordings
- Software
- Spreadsheets / CSV files
- Surveys
- Technical reports
- Teaching tools designed by faculty
  - Theses
- Transcripts
- Video
- Visualizations
- Websites
- \Mhita papara



# Data Management vs.

# Data Management Plans



# Data Management vs DMPs

- Data management refers to the things researchers do as they create, collect, describe, store, and work with data generated by their research.
- Data Management Plans (DMP) are a written description of what data is generated or otherwise acquired for a particular research project, and how the data will be utilized and stored during and after a project.

# Data Management vs DMPs

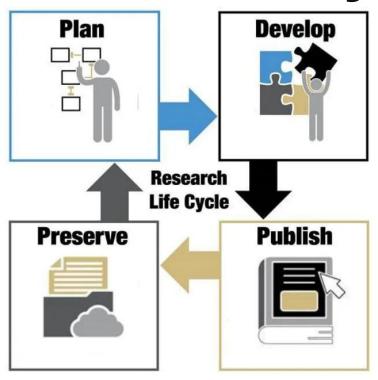
- Data management is what you do.
- Data Management Plans are where you write about what you're going to do. (Usually because a grant requires one.)

#### Why is data management important?

- Protect data from loss
- Saves you time
- You can find your data when you need it
- Helps new members of research teams understand processes faster
- Facilitates reproducibility
- Improves the quality of published data
- Keep sensitive data secure



# Research Lifecycle



# Data Storage & Access



## **Data Storage & Access**

- Where will the data be stored?
  - Data being collected
  - Data being analyzed
- Who can access the data?
  - How can they access it?
  - Is there sensitive or confidential data?
  - What security measures are in place to protect the data?

# Things to think about

- Who's paying for data storage?
- How long will you have access to this storage?
  - What happens when you graduate?

# **Big Data**

- Larger data sets means increased complexity!
- It's harder to store and provide access to data when you're working with terabytes or petabytes of data

# Backups & Versioning



# Why backup your data?

- Technology failure
- Natural disasters
- Theft
- Human error

#### 3-2-1 Rule

- Three copies
  - One primary and two backups
- Two formats/media
  - e.g. External hard drive & cloud storage
- One off site
  - Where is your cloud storage located?

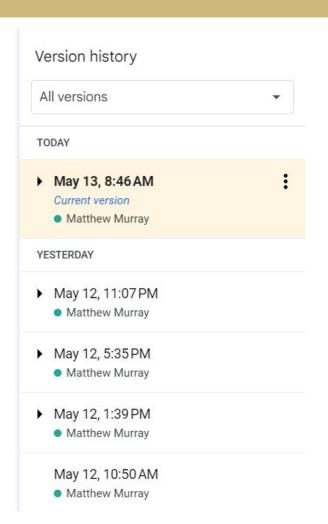
## **USBs are Not Data Storage**

You will lose them



# Versioning

- Versioning is when you save specific versions of your files
- Some software does this automatically
- Can be as basic as having "raw" and "cleaned" versions



# **Documenting Data**



#### What is Documentation?

- Documentation is capturing your research process (from data collection through analysis)
- Includes the Five Ws (and One H)
  - What was done, Who did it, When it happened, Where it happened, Why it was done, and How it was done
- Many different ways to do this

# Why is it Important?

- You can keep track of what still needs to be done.
- Your future self will thank you for writing down what you did.
- Allows others to understand your process and replicate your work.

# Why is it Important?

▲ A B	С	D	E F	G	Н	1	J	K L	M	N	0	P	Q	R	S	Т	UV	W	X	Y Z	AA	AB	AC	AD
1 79c7a2ce89 JC199149		03/25/201	30 TRAFFIC	S FUNCTION	CLEAR	DAYLIGHT	TURNING	ONE-WAY	4 STRAIGI	HT DRY	NO DEFEC ON	SCENE	INJURY ANY				OVER \$1,503/25/201	IMPROPE	DRIVING	536 W	RANDOLP	121		
2 792b539dea JB422857		09/05/201	30 NO CON	TF NO CONTE	CLEAR	DAYLIGHT	ANGLE	NOT DIVIE	2 STRAIGI	HT DRY	NO DEFEC NO	T ON S	NO INJURY	DRIVE AW	AY		OVER \$1,509/05/201	VISION OF	FAILING T	4821 N	ELSTON A	1712		
3 0115ade9a7JF318029		07/15/202	30 UNKNOV	W UNKNOW	CLEAR	DARKNESS	ANGLE	NOT DIVIDED	STRAIG	HT DRY	UNKNOW! ON	SCENE	NO INJURY	DRIVE AW	AY		OVER \$1,507/15/202	UNABLE T	UNABLE T	300 N	CENTRAL	1123		
4 017040c619 JF324552		07/15/202	30 TRAFFIC	S FUNCTION	CLEAR	DAYLIGHT	REAR EN	NOT DIVIDED	STRAIGI	HT DRY	NO DEFEC NO	T ON S	NO INJURY	DRIVE AW	AY		OVER \$1,5 07/20/202	UNABLE T	UNABLE T	2359 N	ASHLAND	1811		
5 78eee027ecJB291672		06/03/201	30 NO CON	TF NO CONTE	CLEAR	UNKNOW	PARKED	ONE-WAY	1 STRAIGI	HT DRY	NO DEFEC ON	SCENE	NO INJURY	DRIVE AW	AY	Y	OVER \$1,506/03/201	<b>UNABLE T</b>	UNABLE T	1630 N	KEELER AV	2534		
6 7943cacbae JF330061		07/24/202	25 NO CON	TF NO CONTE	CLEAR	DAYLIGHT	PARKED	VONE-WAY	STRAIG	HT DRY	NO DEFEC ON	SCENE	NO INJURY	DRIVE AW	AY		OVER \$1,507/24/202	NOT APPL	NOT APPL	2817 W	61ST ST	823		
7 01aaa759c6 JF319819		07/15/202	40 NO CON	TF NO CONTE	CLOUDY	C DAYLIGHT	ANGLE	DIVIDED - W/M	EDIA STRAIGI	HT DRY	NO DEFEC NO	T ON S	NO INJURY	DRIVE AW	AY		OVER \$1,5 07/16/202	IMPROPE	NOT APPL	5100 N	LAKE SHO	2024		
8 7b1537e0a3JA252488		05/05/201	30 TRAFFIC	S FUNCTION	CLEAR	DAYLIGHT	SIDESWII	NOT DIVIDED	STRAIGI	IT DRY	NO DEFEC NO	T ON S	NO INJURY	DRIVE AW	AY		\$501 - \$1, 05/05/201	<b>UNABLE T</b>	UNABLET	4498 W	ADDISON	1731		
9 7b4700df14 JB529385		11/26/201	30 TRAFFIC	S FUNCTION	SNOW	DAYLIGHT	ANGLE	NOT DIVIE	2 STRAIGI	HT SNOW C	R NO DEFECTS		NO INJUR'Y				OVER \$1,5 11/26/201	WEATHER	RUNABLE T	8300 S	STONY ISL	412		
10 7b3545fb91 JD298172		07/15/202	25 TRAFFIC	S FUNCTION	CLEAR	DAYLIGHT	PEDALCY	NOT DIVIDED	STRAIG	HT DRY	NO DEFEC ON	SCENE	INJURY ANY				\$500 OR L 07/15/202	DISREGAR	DRIVING	298 S	CICERO AV	1533		
11 79cac9c36c JA441727		09/13/201	30 NO CON	TF NO CONTE	CLEAR	DAYLIGHT	PARKED	PARKING LOT	STRAIG	HT DRY	NO DEFEC NO	T ON S	NO INJURY	DRIVE AW	AY	Υ	OVER \$1,509/22/201	UNABLE T	UNABLE T	225 W	DIVISION	1823		
12 01d4adbecb JF399117		09/16/202	35 NO CON	TF NO CONTE	CLEAR	DARKNESS	REAR TO	FALLEY	STRAIG	HT DRY	NO DEFEC NO	T ON S	INJURY AND	O / OR TOW	DUE TO	Y	OVER \$1,5 09/17/202	UNABLE T	UNABLE T	609 N	CLARK ST	1832		
13 0362b3f8dd JF318271		07/15/202	30 NO CON	TF NO CONTE	RAIN	DAYLIGHT	REAR EN	DIVIDED - W/M	EDIA STRAIGI	HT WET	NO DEFEC NO	T ON S	NO INJURY	DRIVE AW	AY		OVER \$1,5 07/15/202	UNABLE T	UNABLE T	3458 N	CICERO AV	1634		
14 03d3679ef4 JF318956		07/15/202	30 STOP SIG	N FUNCTION	RAIN	DAYLIGHT	PEDESTR	FOUR WAY	STRAIG	HT WET	NO DEFEC ON	SCENE	INJURY ANY			Υ	\$500 OR L 07/15/202	FAILING T	UNABLE T	1900 S	LAFLIN ST	1235		
15 049099b67aJF318460		07/15/202	35 TRAFFIC	S NOT FUNC	RAIN	DAYLIGHT	TURNING	NOT DIVIDED	STRAIG	HT WET	UNKNOW! ON	SCENE	NO INJURY	DRIVE AW	AY		OVER \$1,5 07/15/202	IMPROPE	DRIVING S	4400 W	47TH ST	815	,	Y
16 04ca492845 JF319187		07/15/202	30 NO CON	TF NO CONTE	RAIN	DAYLIGHT	PARKED	NOT DIVIDED	STRAIG	HT WET	NO DEFEC NO	T ON S	NO INJURY	DRIVE AW	AY		\$500 OR L 07/16/202	NOT APPL	NOT APPL	318 S	FEDERAL S	113	,	Υ
17 051542719€ JF319114		07/15/202	30 UNKNOV	W UNKNOW	RAIN	DARKNESS	REAR TO	FDIVIDED - W/M	EDIA STRAIGI	IT WET	UNKNOW! ON	SCENE	INJURY AND	OR TOW	DUE TO	CRASH	OVER \$1,5 07/15/202	NOT APPL	NOT APPL	401 N	WABASH /	1834	,	Y
18 7b1c666d9dJB163972		02/24/201	25 NO CON	TENO CONTE	CLEAR	DAYLIGHT	SIDESWII	DIVIDED -	4 STRAIGI	HT DRY	NO DEFEC NO	T ON S	NO INJURY	DRIVE AW	AY	N	\$501 - \$1, 02/24/201	UNABLE T	UNABLE T	520 S	MICHIGAN	123		
19 79704e1b74JC315780		06/21/201	30 TRAFFIC	SFUNCTION	CLEAR	DAYLIGHT	TURNING	DIVIDED - W/M	EDIA STRAIGI	HT DRY	NO DEFEC ON	SCENE	INJURY AND	O / OR TOW	DUE TO	Y	OVER \$1,506/21/201	FOLLOWIN	UNABLE T	1198 S	STATE ST	123		
20 7994103871JA425956		09/08/201	30 STOP SIG	N FUNCTION	CLEAR	DAYLIGHT	REAR EN	NOT DIVIDED	STRAIGI	TDRY	NO DEFECTS		NO INJUR'Y			Υ	OVER \$1,5 09/10/201	FOLLOWIN	NOT APPL	3003 W	GRANVILL	2413		
21 7b233fb34d JC174024		03/05/201	25 STOP SIG	N FUNCTION	CLEAR	DAYLIGHT	REAR EN	NOT DIVIE	4 STRAIGI	HT SNOW C	R NO DEFEC ON	SCENE	NO INJUR'N	1			\$500 OR L 03/05/201	NOT APPL	UNABLE T	5331 W	FULLERTO	2515		
22 7b3850c200 JC233284		04/22/201	30 NO CON	TF UNKNOW	CLEAR	DARKNESS	FIXED OF	J DIVIDED - W/M	EDIA STRAIGI	T DRY	NO DEFEC ON	SCENE	INJURY AND	OR TOW	DUE TO	CRASH	OVER \$1,5 04/22/201	UNDER TH	UNABLE T	804 S	ASHLAND	1231		
23 054e9bad1eJF318760		07/15/202	30 NO CON	TF NO CONTE	RAIN	DAYLIGHT	PARKED	NOT DIVIDED	STRAIG	HT WET	NO DEFEC ON	SCENE	NO INJURY	DRIVE AW	AY	Y	OVER \$1,5 07/15/202	UNABLE T	UNABLE T	7804 S	KENWOOL	411		
24 056c477498 JF318680		07/15/202	30 NO CON	TENO CONTE	RAIN	DAYLIGHT	REAR EN	NOT DIVIDED	STRAIG	HT WET	UNKNOWION	SCENE	NO INJURY	DRIVE AW	AY	Y	OVER \$1,5 07/15/202	FOLLOWIN	NUNABLE T	3047 W	CERMAK F	1033		
25 05c1521b06 JF318600		07/15/202	30 UNKNOV	WUNKNOW	CLEAR	DAYLIGHT	ANGLE	T-INTERSECTION	N STRAIGI	HT DRY	UNKNOW! NO	T ON S	NO INJURY	DRIVE AW	AY		\$501 - \$1, 07/15/202	NOT APPL	NOT APPL	939 W	18TH ST	1235		
26 05c296dfb5(JF318941		07/15/202	30 UNKNOV	W UNKNOW	FREEZIN	G DAYLIGHT	FIXED OF	J FOUR WAY	STRAIGI	HT WET	NO DEFEC ON	SCENE	INJURY ANY				OVER \$1,5 07/15/202	UNABLE T	NOT APPL	1502 N	CICERO AV	2533		
27 7951cf0b6a; JA562045		12/24/201	30 UNKNOV	W UNKNOW	SNOW	DAWN	REAR EN	NOT DIVIE	2 STRAIGI	HT SNOW C	R UNKNOW! NO	T ON S	NO INJURY	DRIVE AW	AY		OVER \$1,5 12/24/201	TURNING	NOT APPL	2797 N	LARAMIE.	2521		
28 7a08f16969 JE233086		05/14/202	30 NO CON	TENO CONTE	CLEAR	DAYLIGHT	PARKED	NOT DIVIDED	STRAIGI	HT UNKNO	VI UNKNOW! NO	T ON S	NO INJURY	DRIVE AW	AY	Y	OVER \$1,5 05/17/202	UNABLE T	UNABLE T	566 S	KILPATRIC	1131		
29 7be4779908 JB506955		11/07/201	30 NO CON	TENO CONTE	CLEAR	DAYLIGHT	PARKED	NOT DIVIE	2 STRAIGI	HT DRY	NO DEFEC NO	T ON S	NO INJURY	DRIVE AW	AY	Y	OVER \$1,5 11/07/201	UNABLE T	UNABLE T	401 W	WRIGHTW	1935		
30 79f6316976 JE122789		01/25/202	10 NO CON	TENO CONTE	CLEAR	DAYLIGHT	REAR TO	FONE-WAY	STRAIG	HT DRY	NO DEFEC NO	T ON S	NO INJURY	DRIVE AW	AY		OVER \$1,5 01/25/202	IMPROPE	UNABLE T	10343 S	PULASKI R	2211		
31 07baf8ac46(JF318925		07/15/202	30 NO CON	TENO CONTE	RAIN	DAYLIGHT	OTHER N	NOT DIVIDED	STRAIG	HT WET	OTHER ON	SCENE	INJURY AND	O / OR TOW	DUE TO	CRASH	\$501 - \$1, 07/15/202	UNABLE T	NOT APPL	1676 W	OGDEN AV	1224		
32 094ef43765 JF318676		07/15/202	30 TRAFFIC	S FUNCTION	RAIN	DAYLIGHT	TURNING	DIVIDED - W/M	EDIA STRAIGI	HT WET	UNKNOW! NO	T ON S	NO INJUR'Y				OVER \$1,5 07/15/202	FAILING T	FAILING T	5600 W	DIVISION	2531		
33 7b63a439b7JB366967		07/27/201	30 TRAFFIC	SFUNCTION	CLEAR	DARKNESS	PEDALCY	C CENTER T	5 STRAIGI	HT UNKNO	VI UNKNOW! ON	SCENE	NO INJURY	DRIVE AW	AY	Y	\$500 OR L 07/27/201	FOLLOWIN	FOLLOWIN	1941 N	WESTERN	1434		
34 0a43552917JF318249		07/15/202	30 STOP SIG	FUNCTION	RAIN	DAYLIGHT	REAR EN	NOT DIVIDED	STRAIG	HT WET	UNKNOWION	SCENE	NO INJUR'Y				OVER \$1,507/15/202	FOLLOWIN	UNABLE T	2857 N	KEDZIE AV	1411		
35 Obe1398662JF319154	N	07/15/202	30 NO CONT	TENO CONTE	RAIN	DARKNESS	TURNING	NOT DIVIDED	STRAIGI	HT WET	NO DEFEC ON	SCENE	INJURY AND	O / OR TOW	DUE TO	Y	\$501 - \$1, 07/15/202	FAILING T	NOT APPL	5335 W	CHICAGO	1524		
36 7a77bed557JE267857		06/16/202	30 TRAFFIC	SFUNCTION	CLEAR	DAYLIGHT	REAR EN	DIVIDED - W/M	EDIA STRAIGI	T DRY	NO DEFEC NO	T ON S	NO INJURY	DRIVE AW	AY		OVER \$1,5 06/16/202	UNABLE T	UNABLE T	1554 W	69TH ST	725		
37 Oda8a78c1aJF318221	Y	07/15/202	30 NO CON	TENO CONTE	CLEAR	UNKNOW	PARKED	NOT DIVIDED	STRAIGI	HT DRY	NO DEFEC NO	T ON S	NO INJURY	DRIVE AW	ΑY	Y	OVER \$1,507/15/202	UNABLE T	NOT APPL	5925 N	NORTHWE	1611		
38 7a98b33a77JB311906		06/16/201	30 TRAFFIC	SFUNCTION	CLEAR	DAYLIGHT	TURNING	NOT DIVIDED	STRAIG	HT DRY	NO DEFEC NO	T ON S	NO INJUR'Y				\$501 - \$1, 06/18/201	FOLLOWIN	UNABLE T	5700 N	KENMORE	2022		

#### **Documentation matters**

- HIT\_AND\_RUN\_I
- STATEMENTS\_TAKEN\_I

# **Examples of Documentation**

- Data Dictionaries
  - Contains a description of elements in a dataset, including names, definitions, acronyms, and other relevant information.
- README files
  - A plain text file that provides information and instructions about a project, including its purpose, usage instructions, known issues, and contact information for support or collaboration.

#### **README Files**

- Title of Dataset
- Authors
- Contact information
- Date of data collection
- Licenses/restrictions placed on the data
- Links to publications that cite or use the data
- Recommended citation for the data
- Structure and organization of the data files
- Descriptions of variables
- List of software and version numbers

#### Metadata

- Data about data!
- Descriptions that help you find and understand data
- Different fields/disciplines use different metadata standards

#### Creator Gifford, Lauren Nacu-Schmidt, Ami Osborne-Gowey, Jeremiah Boykoff, Max **Date Issued** 2023-04 **Academic Affiliation** Cooperative Institute for Research in Environmental Sciences **Last Modified** 2023-05-02 Resource Type Data Set **Rights Statement** In Copyright 🖸 DOI ttps://doi.org/10.25810/c862-0e81.60 Language English [eng] License Creative Commons BY Attribution 4.0 International

# File Names & Structures



#### "FINAL"doc







FINAL rev. 2 doc







FINAL\_rev.6.COMMENTS.doc

FINAL\_rev.8.comments5. CORRECTIONS. doc







FINAL\_rev.18.comments7. corrections9.MORE.30.doc

FINAL\_rev.22.comments49. corrections 10. #@\$%WHYDID ICOMETOGRADSCHOOL?????.doc

PHD Comics: NotFinal.Doc. Jorge Cham, 2012. https://phdcomics.com/ comics.php?f=1531.

> xkcd: Documents. Randall Munroe. https://xkcd.com/1459/



PROTIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.

#### **Bad File Names**

- Traffic\_Crashes\_-\_Crashes.csv
- "Traffic Crashes Crashes."
   data.cityofchicago.org, 2023-05-04.
   https://catalog.data.gov/dataset/traffic-crashe s-crashes.

# File Naming Best Practices

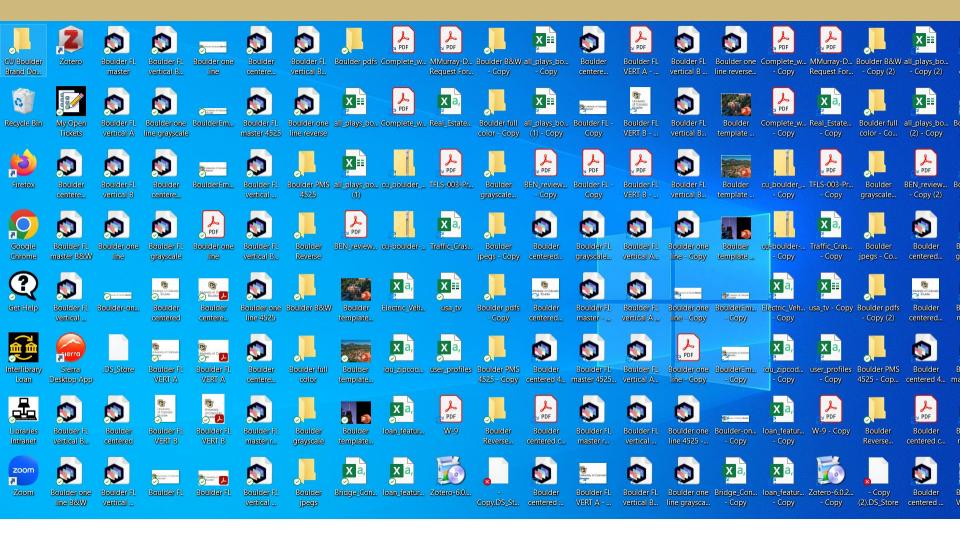
- Be consistent
  - Do: Use the same format for all files
  - Don't: Keep changing file names
- Be descriptive
  - Do: Avoid generic terms
  - Don't: Use "Final" in your file name
- Limit file name length
  - Do: Use abbreviations
  - Don't: write-out-every-word-in-your-data-file.xlsx

# File Naming Best Practices

- Use CamelCase and dashes
  - Do: FileName-2023.pdf
  - Don't: use spaces inyourfilenames.doc
- Use standardized numbers and dates
  - Do: Use leading zeros (001.png)
  - Don't: Have files named 9-12-11.csv
- Use the Latin alphabet
  - Don't: Use punctuation or special characters

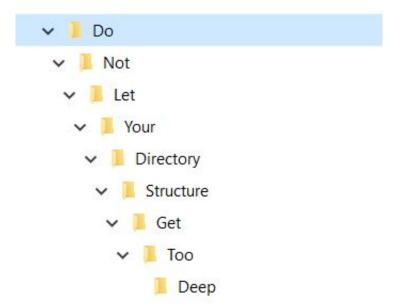
#### **Dates**

- ISO 8601
- 2023-05-13
- Remember that historically dates were not consistent
  - 1712-02-30 (February 30th, 1712) was a real date in Sweden



## File Structuring

- Don't save everything to the desktop
- Don't let your structure get too deep



# File Formats & Units of Measurement



#### **File Formats**

- Non-proprietary/Open
  - Can still be used even if original software is inaccessible
  - e.g. Use CSV files instead of .vc
- Unencrypted
- Uncompressed/lossless

#### **Units of Measurement**

- Be consistent
- Use standardized measurements
- Don't keep changing between them
- How big is a "cup"?
  - US Cup: 8 oz (236.6 ml)
  - Metric cup: 250 ml
  - Japanese rice cup: 180 ml

# **Data Preservation**



#### What is Data Preservation?

- What you do with your data after the research project is over
- Not all data has to be preserved!
  - What data should be preserved?
  - What data can be gotten rid of? When do you get rid of it?
- All the stuff we talked about before (storage, access, metadata, etc.)

# Data Management Plans



#### What is a DMP?

- A framework for how you'll manage your data
- Describe your plans for collecting, organizing, storing, and sharing your data
- About two pages long

## When should you make a DMP?

- When you have to for grant requirements
  - Different grants have different requirements
- When you don't have to, but want to, keep track of your data

#### Two types of DMPs

- The "final" version you'll submit with a grant proposal
- The "living" version that you'll continue to update as your research project progresses

#### What goes in a DMP?

- Data generated in this project
- Software and file formats that will be used
- Where data will be stored and who can access it
- Any privacy, legal, or ethical constraints
- Metadata standards
- How the data will be preserved/shared
- How the data can be reused

# **DMP Tool**



#### **DMPTool**

- Has pre-formatted DMP templates from various funding agencies
- Walks you through the process of completing the DMP

### **DMPTool Activity**

- Go to <a href="https://dmptool.org/">https://dmptool.org/</a>
- When signing in, indicate that you're from CU Boulder
- Select a DMP Template relevant for your field (if you're having trouble choosing, pick the generic NSF template)
- Look through the prompts
- Attempt to answer the prompts in the context of your proposed project

## **DMPTool Activity**

- Were any of the prompts challenging or confusing?
- Is there anything you would like clarification on?

 We're happy to read drafts and provide feedback on your draft DMPs, so please send them our way!

# Other Useful Tools



## **Data Management Tools**

- Git and GitHub
  - Track changes to data and code
- Open Science Framework
  - One-stop shop for project management
- Open Refine
  - Clean data
- File Renamers (various)
  - Rename files so they're consistent
- Zotero
  - Reference management software

## Consultations



# How CRDDS can assist with Data Management

- One-on-one or small group consults
- Review draft DMPs and README files
- Help navigate data policies
- Find data repositories
- Advice on file formats, etc.
- Email us: <a href="mailto:crdds@colorado.edu">crdds@colorado.edu</a>