University of Idaho
Library

MAKE DATA MANAGEMENT EASIER

JEREMY KENYON
RESEARCH LIBRARIAN
jkenyon@uidaho.edu

Summary

- Introduction
- Data Hub
- Tips for Data Management
- Q&A

Data Hub: Geospatial and Data Sciences Support

- U of I Library Data Hub
  - Located in the Map Room, First Floor, Rm 107
  - Individual workstations for specific research software and tools
  - Collaborative work areas focused on supporting data sciences analysis and visualization
  - Service desk staffed 11am-3pm, M-F by U of I Data and GIS Librarians, and others
  - Other Campus Units encouraged to collocate in the Data Hub, including RCDS and Statistical Consulting
  - Website: https://www.lib.uidaho.edu/datahub/

Tip #1: Backup your Data

Common problems

- Corrupted data, failed hard drive, laptop lost/stolen, mistakes (deletions, user error)

3-2-1 Rule

- Have at least 3 copies of your data
- Store them in 2 different media
- Keep 1 copy off-site (geographically differentiated)

Example plan:

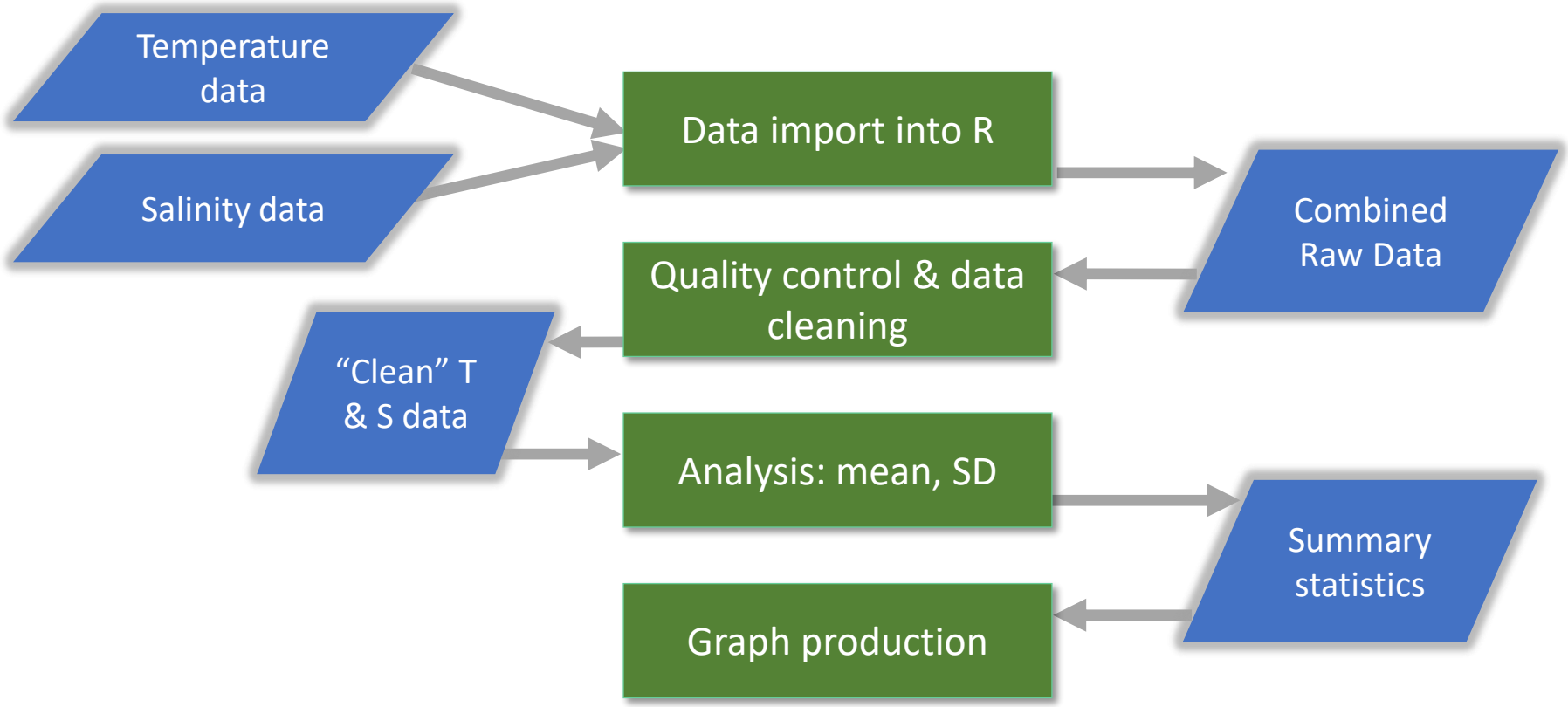
- One copy on local hard drive
- One copy on OneDrive (geographic replication off-site)
- One copy on a physical media device

Image from: Vera. 2020. “Best Practice: 3-2-1 Backup Strategy for Home Users & Businesses [Clone Disk].” Retrieved from: https://www.partitionwizard.com/clone-disk/backup-strategy.html

Also adapted from: University of Virginia Library Research Data Services + Sciences. 2020. “Data Storage and Backups.” Retrieved from: https://data.library.virginia.edu/data-management/plan/storage/

Tip #2: Never modify raw data; *version* it as you go

As you work on data, copy it and modify the copy, saving it as a new file. Do so repeatedly to avoid changing the original data. If desired, just write-protect the original data.



From: DataONE. 2012. "DataONE Education Module: Analysis and Workflows." Retrieved from: http://www.dataone.org/sites/all/documents/L10_Analysis Workflows.pptx

Tip #2: Never modify raw data; version it as you go

One recommended procedure is to simply copy your entire project folder (excepting large data files) periodically to maintain old versions.

```
.
|--project_name
|    |--current
|    |    |--...project content as described earlier...
|    |--2016-03-01
|    |    |--...content of 'current' on Mar 1, 2016
|    |--2016-02-19
|    |    |--...content of 'current' on Feb 19, 2016
```

From: Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK (2017) Good enough practices in scientific computing. PLoS Comput Biol 13(6): e1005510. https://doi.org/10.1371/journal.pcbi.1005510

Tip #3: Use clear, unambiguous data values when possible

Consider the purpose of entering something in a cell or column. Identifiers can be the means to re-using the data with other datasets in the future.

Two types of identifiers:

- Standardized identifiers: ISBNs, DOIs, species names, language codes
  - Usually a part of an international registry system
- Localized identifiers: record IDs, database keys, site/plot codes, other enumerated codes

index	issn-elec	affiliation	rank	DOI	link	container-title	issued	
Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filte
0	1477-0296	University of Idaho, USA	1	10.1177/030913331988...	http://journal...	Progress in Physical Ge...	2019-11-19	Univ
1	2574-0962	Department of Chemistry, University ...	1	10.1021/acsaem.9b01889	https://pubs.a...	ACS Applied Energy Mat...	2019-11-18	Offic
2	1365-2664	Department of Biological Sciences U...	1	10.1111/1365-2664.13539	https://api.wil...	Journal of Applied Ecology	2019-11-16	NUL
3	1557-8070	Department of Physics, University of ...	1	10.1089/ast.2018.1972	https://www.l...	Astrobiology	2019-11-15	NUL
4	1574-6941	Department of Molecular & Cell Biolo...	1	10.1093/femsec/fiz182	http://acade...	FEMS Microbiology Ecol...	2019-11-15	NUL
5	1996-8175	Research Associate, Academy of Nat...	1	10.1002/tax.12124	https://api.wil...	TAXON	2019-11-15	NUL
6	1788-9170	State Key Laboratory of Crop Biology...	1	10.1556/0806.47.2019.44	https://www....	Cereal Research Comm...	2019-12	NUL

Tip #3: Use clear, unambiguous data values when possible

Also, think about your variable names

Good Name	Good Alternative	Avoid
Max_temp_C	MaxTemp	Maximum Temp (°C)
Precipitation_mm	Precipitation	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell Type
Observation_01	first_observation	1st Obs

From: Hoyt et al. (2019, July 5). datacarpentry/spreadsheet-ecology-lesson: Data Carpentry: Data Organization in Spreadsheets for Ecologists, June 2019 (Version v2019.06.2). Zenodo. http://doi.org/10.5281/zenodo.3269869

Tip #3: Use clear, unambiguous data values when possible

- In many cases, zero *is* a value, it means something for an observation to be recorded as zero.

- In other cases, you simply don't have a value. Don't use zero here, but don't use nothing either. Pick an unrealistic value for your data, like -999, or a code like "NA" or "NULL".

- Alternatively, consider error codes (e.g. -333) for cases where you need to note something other than "no value"

Tip #4: Create metadata

DLC Animal List variable descriptions

count	Animal List variable name	Animal List Variable Definition
1	Taxon	Taxonomic code: In most cases, comprised of the first letter of the genus and the first three letters of the species; if taxonomic designation is a subspecies, comprised of the first letter of genus, species and subspecies, and hybrids are indicated by the first three letters of the genus. See Table 1 for details.
2	DLC_ID	Specimen ID: Unique identification number assigned by the DLC at accession of animal.
3	Hybrid	Hybrid status: N=not a hybrid. S=species hybrid. B=subspecies hybrid. If sire is one of multiple possible and animal could be a hybrid, it is designated a hybrid.
4	Sex	Sex: M=male. F=Female. ND=Not determined
5	Name	House name: Animal name assigned at DLC
6	Current_Resident	Resident status: Whether or not the animal currently lives in the DLC colony.

https://dx.doi.org/10.1038%2Fsdata.2014.19

Tip #4: Create metadata

Potential Fields to Include:

- Variable Name
- Variable Definition
- Variable Definition Source
- How measured
- Data units
- Data format
- Min/max values
- Coded values/defs
- Null values representation
- Precision of measurement
- Known issues
- Relationship to other variables
- Other notes

	A	B	C	D
1	name	plot_name	group	description
2	mouse	Mouse	demographic	Animal identifier
3	sex	Sex	demographic	Male (M) or Female (F)
4	sac_date	Date of sac	demographic	Date mouse was sacrificed
5	partial_inflation	Partial inflation	clinical	Indicates if mouse showed partial pancreatic inflation
6	coat_color	Coat color	demographic	Coat color, by visual inspection

From: Broman & Woo. (2018) Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, DOI: 10.1080/00031305.2017.1375989

Tip #5: Pick the right tabular format

Location	2017	2018	2019
Moscow	32	15	98
Coeur d'Alene	74	38	105
Boise	143	67	192

Wide data: good for human consumption, final outputs for people to read

Year	Location	Count
2017	Moscow	32
2017	Coeur d'Alene	74
2017	Boise	143
2018	Moscow	15
2018	Coeur d'Alene	38
2018	Boise	67
2019	Moscow	98

Long or "Tidy" data:
good for machine consumption, for analysis or visualization

For more on 'tidy' data: Wickham & Grolemund. (2017). "Tidy data." *R for Data Science.* https://r4ds.had.co.nz/tidy-data.html

Tip #5: Pick the right tabular format

Using scripting tools like R and Python, flipping back and forth is relatively feasible.

- R (using tidyr): gather() and spread()
- Python (using pandas): pivot() and melt()

Other tools, e.g. SPSS/SAS/Stata/Tableau/Oracle Analytics, possess features for reshaping data too.

Tip #6: Use standardized date time formats

Common examples of date and times:

	A	B	C	D	E	F	G	H	I
1	What I typed in	day-month	DOW, month, day, year	month-year	Initial-year	M/D/YYYY	DD/MM/YYYY	DD/MM/YY	number
2	2-jul	2-Jul	Wednesday, July 02, 2014	Jul-14	J-14	7/2/2014	02/07/2014	07/02/14	41822
3	Jul-14	14-Jul	Monday, July 14, 2014	Jul-14	J-14	7/14/2014	14/07/2014	07/14/14	41834
4	1-jan-1900	1-Jan	Sunday, January 01, 1900	Jan-00	J-00	1/1/1900	01/01/1900	01/01/00	1

The problem, beyond inconsistency, is that systems may not know how to read the string.

From: https://datacarpentry.org/spreadsheets-socialsci/03-dates-as-data/index.html

Tip #6: Use standardized date time formats

The international standard for displaying date and times is codified in ISO 8601.

```
   Year:
      YYYY (eg 1997)
   Year and month:
      YYYY-MM (eg 1997-07)
   Complete date:
      YYYY-MM-DD (eg 1997-07-16)
   Complete date plus hours and minutes:
      YYYY-MM-DDThh:mmTZD (eg 1997-07-16T19:20+01:00)
   Complete date plus hours, minutes and seconds:
      YYYY-MM-DDThh:mm:ssTZD (eg 1997-07-16T19:20:30+01:00)
   Complete date plus hours, minutes, seconds and a decimal fraction of a
second
      YYYY-MM-DDThh:mm:ss.sTZD (eg 1997-07-16T19:20:30.45+01:00)
```

From: Wolf & Wicksteed. (1997). *Date and Time Formats.* https://www.w3.org/TR/NOTE-datetime

Tools are built to understand this format. Often, they enable derivative data to be produced, like month or day of the week.

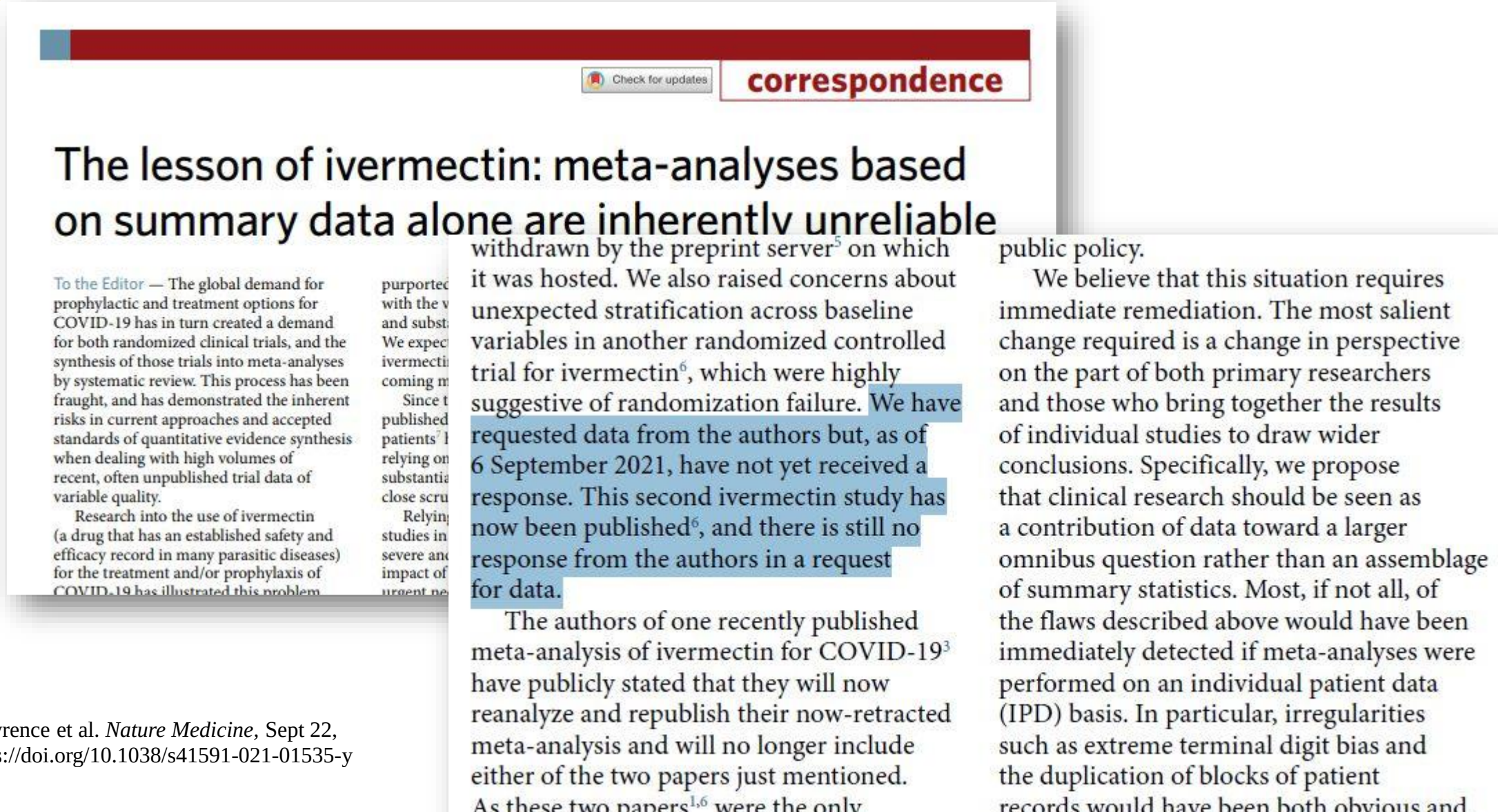
Tip #7: Assume others will see your data.

- Data publishing, sharing, reproducibility, open science. All introduce reasons for people to see your data.

- Remember:
    - Most funders require data sharing.
    - Many journals expect data sharing.

- Reduce fear or anxiety about others viewing your work by maintaining good practices (or good enough) during your data management.

Tip #7: Assume others will see your data.



Check for updates

correspondence

The lesson of ivermectin: meta-analyses based on summary data alone are inherently unreliable

To the Editor — The global demand for prophylactic and treatment options for COVID-19 has in turn created a demand for both randomized clinical trials, and the synthesis of those trials into meta-analyses by systematic review. This process has been fraught, and has demonstrated the inherent risks in current approaches and accepted standards of quantitative evidence synthesis when dealing with high volumes of recent, often unpublished trial data of variable quality.

Research into the use of ivermectin (a drug that has an established safety and efficacy record in many parasitic diseases) for the treatment and/or prophylaxis of COVID-19 has illustrated this problem

withdrawn by the preprint server[5] on which it was hosted. We also raised concerns about unexpected stratification across baseline variables in another randomized controlled trial for ivermectin[6], which were highly suggestive of randomization failure. We have requested data from the authors but, as of 6 September 2021, have not yet received a response. This second ivermectin study has now been published[6], and there is still no response from the authors in a request for data.

The authors of one recently published meta-analysis of ivermectin for COVID-19[3] have publicly stated that they will now reanalyze and republish their now-retracted meta-analysis and will no longer include either of the two papers just mentioned. As these two papers[1,6] were the only

public policy.

We believe that this situation requires immediate remediation. The most salient change required is a change in perspective on the part of both primary researchers and those who bring together the results of individual studies to draw wider conclusions. Specifically, we propose that clinical research should be seen as a contribution of data toward a larger omnibus question rather than an assemblage of summary statistics. Most, if not all, of the flaws described above would have been immediately detected if meta-analyses were performed on an individual patient data (IPD) basis. In particular, irregularities such as extreme terminal digit bias and the duplication of blocks of patient records would have been both obvious and

From: Lawrence et al. *Nature Medicine*, Sept 22, 2021. https://doi.org/10.1038/s41591-021-01535-y

Summary

1. Backup up your data using the 3-2-1 rule.
2. Never modify raw data; version your data.
3. Be intentional with data values.
4. Use a data dictionary or codebook (at least!).
5. Be aware of long vs. wide data formats.
6. Use standardized data time formats.
7. Assume others will see your data and act accordingly.

Fall 2023 Graduate Student Essentials

When: Wednesdays from 12:30pm – 1:30pm
Where: Library first floor classroom (Room 120) and live via Zoom

~~**September 6**: Essential Library Resources, Services, and Skills for Graduate Students~~

~~**September 13**: Getting started with Zotero~~

~~**September 20**: The Textbooks are Too Expensive! Open Educational Resources and Graduate Students~~

~~**September 27**: Six Questions to Ask before Publishing Your Journal Article~~

October 4: Make Data Management Easier

October 11: Web Mapping for Every Discipline – How to Use ArcGIS Online