7 Tips To Make Your Data Management Life Easier

Jeremy Kenyon
Research Librarian

jkenyon@uidaho.edu
Summary

• Introduction
• Data Hub
• 7 Tips for Data Management
• Q&A
Data Hub: Geospatial and Data Sciences Support

• U of I Library Data Hub
  o Open now!
  o Located in the Map Room, First Floor, Rm 107
  o Individual workstations for specific research software and tools
  o Collaborative work areas focused on supporting data sciences analysis and visualization
  o Service desk staffed 11am-3pm, M-F by U of I Data and GIS Librarians, and others
  o Other Campus Units encouraged to collocate in the Data Hub, including RCDS and Statistical Consulting
  o 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

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.

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
```

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

Also, think about your variable names

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

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: Use a data dictionary or codebook

<table>
<thead>
<tr>
<th>count</th>
<th>Animal List variable name</th>
<th>Animal List Variable Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Taxon</td>
<td>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.</td>
</tr>
<tr>
<td>2</td>
<td>DLC_ID</td>
<td>Specimen ID: Unique identification number assigned by the DLC at accession of animal.</td>
</tr>
<tr>
<td>3</td>
<td>Hybrid</td>
<td>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.</td>
</tr>
<tr>
<td>4</td>
<td>Sex</td>
<td>Sex: M=male. F=Female. ND=Not determined</td>
</tr>
<tr>
<td>5</td>
<td>Name</td>
<td>House name: Animal name assigned at DLC</td>
</tr>
<tr>
<td>6</td>
<td>Current_Resident</td>
<td>Resident status: Whether or not the animal currently lives in the DLC colony.</td>
</tr>
</tbody>
</table>

https://dx.doi.org/10.1038%2Fsdata.2014.19
Tip #4: Use a data dictionary or codebook

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

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>name</td>
<td>group</td>
<td>description</td>
</tr>
<tr>
<td>2</td>
<td>mouse</td>
<td>Mouse</td>
<td>Animal identifier</td>
</tr>
<tr>
<td>3</td>
<td>sex</td>
<td>Sex</td>
<td>Male (M) or Female (F)</td>
</tr>
<tr>
<td>4</td>
<td>sac_date</td>
<td>Date of sac</td>
<td>Date mouse was sacrificed</td>
</tr>
<tr>
<td>5</td>
<td>partial_inflation</td>
<td>Partial inflation</td>
<td>Indicates if mouse showed partial pancreatic inflation</td>
</tr>
<tr>
<td>6</td>
<td>coat_color</td>
<td>Coat color</td>
<td>Coat color, by visual inspection</td>
</tr>
</tbody>
</table>

Tip #5: Pick the right tabular format

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

<table>
<thead>
<tr>
<th>Location</th>
<th>2017</th>
<th>2018</th>
<th>2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moscow</td>
<td>32</td>
<td>15</td>
<td>98</td>
</tr>
<tr>
<td>Coeur d'Alene</td>
<td>74</td>
<td>38</td>
<td>105</td>
</tr>
<tr>
<td>Boise</td>
<td>143</td>
<td>67</td>
<td>192</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Year</th>
<th>Location</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017</td>
<td>Moscow</td>
<td>32</td>
</tr>
<tr>
<td>2017</td>
<td>Coeur d’Alene</td>
<td>74</td>
</tr>
<tr>
<td>2017</td>
<td>Boise</td>
<td>143</td>
</tr>
<tr>
<td>2018</td>
<td>Moscow</td>
<td>15</td>
</tr>
<tr>
<td>2018</td>
<td>Coeur d’Alene</td>
<td>38</td>
</tr>
<tr>
<td>2018</td>
<td>Boise</td>
<td>67</td>
</tr>
<tr>
<td>2019</td>
<td>Moscow</td>
<td>98</td>
</tr>
</tbody>
</table>

Tip #5: Pick the right tabular format

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

- R (using tidyverse):
  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:

<table>
<thead>
<tr>
<th>What I typed in</th>
<th>day-month</th>
<th>DOW, month, day, year</th>
<th>month-year</th>
<th>Initial-year</th>
<th>M/D/YYYY</th>
<th>DD/MM/YYDD/MM/YYYY</th>
<th>DD/MM/YY</th>
<th>number</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-jul</td>
<td>2-Jul</td>
<td>Wednesday, July 02, 2014</td>
<td>Jul-14</td>
<td>J-14</td>
<td>7/2/2014</td>
<td>02/07/2014</td>
<td>07/02/14</td>
<td>41822</td>
</tr>
<tr>
<td>Jul-14</td>
<td>14-Jul</td>
<td>Monday, July 14, 2014</td>
<td>Jul-14</td>
<td>J-14</td>
<td>7/14/2014</td>
<td>14/07/2014</td>
<td>07/14/14</td>
<td>41834</td>
</tr>
<tr>
<td>1-jan-1900</td>
<td>1-Jan</td>
<td>Sunday, January 01, 1900</td>
<td>Jan-00</td>
<td>J-00</td>
<td>1/1/1900</td>
<td>01/01/1900</td>
<td>01/01/00</td>
<td>1</td>
</tr>
</tbody>
</table>

The problem, beyond inconsistency, is that systems may not know how to read the string.
Tip #6: Use standardized date time formats

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

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.

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 purporting to disagree with the vast majority of the data and the results and conclusions of our analyses. We expect that our results will have an important impact on future research and 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

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 2022 Graduate Student Essentials

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

September 6: Essential Library Skills to Ace Graduate School
September 13: 7 Tips To Make Your Data Management Life Easier
September 20: 3 Simple Tips for Expanding Your Literature Review
September 27: Supercharging Your Scholarly Presence in 3 Easy Steps
October 4: Web Mapping for Every Discipline – How to Use ArcGIS Online