THE DATA PIPELINE:
From Design to Dissemination
THE PURPOSE OF THIS GUIDE

Comply with mandates | Develop robust day-to-day data practices | Access campus resources and expertise
This guide was developed by faculty and staff at MSU to provide guidance to MSU researchers interested in robust data management, collection, cleaning, storage and publication practices. Original contributors included Sara Mannheimer, Data Librarian at MSU Library; Lillian S. Lin, Ph.D., Director of Statistical Consulting and Research Services, Department of Mathematical Sciences; Robin Clausen, Ph.D., Project Manager, and Eric Raile, Ph.D., Director of the Human Ecology Learning & Problem Solving (HELPS) Lab; Aurelien Mazurie, Ph.D., Director of the Bioinformatics Core, Research Cyberinfrastructure; and Micaela Young, Pre-Award Specialist, Office of Sponsored Programs. The guide was updated by Sara Mannheimer in September 2019.

This guide will help MSU researchers with the following:

**Comply with mandates**
The federal government and many private funders have established mandates that require research data to be made accessible to the public. Additionally, some publishers—like PLOS—require publication of supporting data as a condition for article publication.

**Develop robust day-to-day data practices**
Robust data practices mean better overall research and reproducibility of results, which benefits the public and scientific community. Good data practices also reduce the workload for grant submissions, reporting, and publication.

**Access campus resources and expertise on the following topics:**
- Data management planning
- Data collection and cleaning
- Data analysis
- Data storage
- Data publication and dissemination
DATA MANAGEMENT PLANNING

Why Manage Data? | Data Management Plans
Why Manage Data?

**Comply with federal mandates.** Many funding agencies require that grant proposals include a plan for how research data will be managed and shared. As of February 2013, non-classified data produced with federal funds must be made public within twelve months of an article’s publication.

**Facilitate reuse.** Sharing data facilitates scientific discourse and discovery. Archived Hubble telescope data is a great example—astronomers have used this data to make new discoveries without spending time and money on new observations (read more from The Astrophysical Journal: HST/NICMOS Detection of HR 8799 b in 1998, available at [http://doi.org/10.1088/0004-637X/694/2/L148](http://doi.org/10.1088/0004-637X/694/2/L148)).

**Make your data meaningful.** Planning ahead, organizing, and documenting your data makes your data comprehensible to you and your team during the research process, and allows others to understand the data once it has been publicly archived.

**Increase the impact of your research.** Studies have shown that open data can lead to increased citations of associated articles (Sharing Detailed Research Data Is Associated with Increased Citation Rate and Data Reuse and the Open Data Citation Advantage).

Data Management Plans

Data management guidelines can vary between funders. Be sure to check the data management planning guidelines for your granting agency or funder. The University of Minnesota keeps an updated list here: [www.lib.umn.edu/datamanagement/funding](http://www.lib.umn.edu/datamanagement/funding).

Data management plans are usually 1–2 page documents that describe how you will manage your data and make it available once the grant ends. Generally, data management plans contain the following information:

- Types of data
- Contextual details (to make data meaningful to others)
- Storage, backup, and security
- Provisions for protection/privacy
- Policies for reuse
- Policies for access and sharing

Data Management Toolkit

A Data Management Toolkit with templates and boilerplate language is available on the library website: [http://lib.montana.edu/services/data/toolkit](http://lib.montana.edu/services/data/toolkit).
DATA COLLECTION AND CLEANING
Types of Research

1. Foundational Research and Early Stage or Exploratory Research contributes to core knowledge.
2. Design and Development Research develops solutions aimed at achieving a goal.
3. Efficacy, Effectiveness and Scale-up Research contributes to the evidence of impact. This research shows how a certain intervention achieves its intended goals.

Research Design
Beginning with a rigorous design is key to any type of quality research. A few strategies contribute to an acceptable research design.

- Begin with the end in mind—what is the primary goal of the research?
- Decide what type(s) of data you want to collect based on your goals. For example:
  - Summary statistics
  - Comparisons between groups
  - Changes in a measure of interest
  - Contrasts between changes
- Design your study based on the desired type(s) of data.
- Plan your analysis based on the design.
- Document your plan for data collection and analysis.
- Calculate sample size based on the analysis plan.
- Decide what indicators to collect based on your analysis plan.
- Design your spreadsheet and codebook based on your research design and the data that will be collected.
Data Collection
When collecting data, it is easy to make small errors like measurement errors that can affect the quality of the research as a whole. Careful data collection strategies can help prevent errors.

Non-human subjects research:
- Physical instruments
- Plants and animals
- Financial records
- Documents
- Secondary data

Human subjects research:
- Interviews
- Group interviews
- Focus groups
- Surveys
- Experiments
- Observation

Methods of data collection
The methods used to collect data depend on the goals of the research and planned analyses.
- Methods should be driven by research questions, literature, theories, hypotheses, and necessary variables.
- Do not collect more data than is necessary. Be comprehensive, but do not ask unnecessary questions.
- Keep human subjects principles in mind, including minimizing harm, securing informed consent, and protecting privacy.
- Keep detailed records.
- There are additional considerations for panel/time series data (e.g. attrition of participants, consistency of measurement).

Avoiding Error in Data Collection
Limiting sources of error will help produce more definitive research results.
- Use appropriate levels of measurement (i.e., nominal, ordinal, interval, ratio).
- Validity and reliability are very important.
- Construction of scales and indexes.
- Using pre-existing indicators from the literature means that they have been accepted and tested by the community.
- Sampling and non-response bias.
- There are many different representative and non-representative sampling methods available.
- Use techniques for limiting non-response bias, such as repeated contact and incentives.

Other sources of error include:
- Problematic or difficult language in questions/items. Use language that is easy for respondents to understand.
- Order of survey questions/items. Strategies to avoid cognitive biases include moving from general to specific questions, and starting with questions that seem relevant to the respondents in order to draw them in.
- Interviewer characteristics and interview dynamics. Be aware of how these factors may influence respondents.
- Characteristics and interpretations of research participants. Be aware of language and cultural issues.
- Differences in setting and administration. Collecting data in a lab can help control variables.
- Data entry, cleaning and processing. Implement systems that provide redundancy and spot checking to reduce error.

Data Cleaning
Careful documentation of procedures improves the clarity of the data collected and improves the quality of data analysis. Researchers should document the recoding of variables, labeling of variables and values, reorganization of spreadsheets, creation of new datasets, etc.
- Detailed codebooks are essential. Good codebooks show variable names, question language, response options, values assigned to response options, labels for variable values, etc.

The HELPS Lab provides:
- Research and data collection design
- Sampling design
- Data collection: interviews, focus groups, surveys, experiments
- Transcription
- Data cleaning

For more information, see http://helpslab.montana.edu
DATA ANALYSIS:
MISCONCEPTIONS AND TRUTHS, PITFALLS AND SOLUTIONS
Data Analysis

Contact
Statistical Consulting and Research Services
Mark Greenwood, Director
406-994-5594 | www.montana.edu/statisticalconsulting

Statistical Methods
Misconception: Many believe that statistical approaches only involve hypothesis testing and modeling.

Truth: Statistical approaches should be followed whenever data are examined. This includes:
• Visualization
• Descriptive Statistics
• Hypothesis Testing
• Modeling
• Interpretation

Pitfall: A common pitfall involves diving right into the data to see whether your hypothesis is proven or not.

Solution: Plan the analysis carefully
• Before data are collected
• While data are being collected
• While data are being cleaned

Pitfall: Making ad hoc decisions that impact the analyses in different ways instead of following the analysis plan.

Solution: Instead, make decisions at a more general level
• For example, treat similar situations similarly. If you conduct separate analyses for men and women, all decisions that are unrelated to gender should be the same for both groups.

• Decide which observations will be in all analyses rather than letting item missingness decide for you. In other words, you may have to create a different response or code that includes missing data.
• You will want to say at the end of the day that you were even-handed and conducted analyses consistently vs. trying to make up for the unexpected.

Plotting Results
Misconception: The “mark one eyeball test” or seeing an obvious difference on a plot.

Truth: Running a statistical test is the only way to determine significant differences and statistical trends.

Misconception: Believing that Excel is the easiest to use for graphical methods.

Truth: Excel is very limited and difficult to produce multiple plots because it was not designed for research per se, but for financial accounting. There are other, better options available.

Pitfall: A bar graph can be used to show averages.

Solution: A bar graph should be used for counts, while a line graph should be used for averages. This is not quite accurate because bar graphs were designed to show you how many, or what percent and they should only be used for this reason. Line graphs are often better.
Descriptive Statistics

*Pitfall:* Jumping straight into analysis.

*Solution:* You need to think through what type of data you have and different types of analysis methods that are available for that data.

*Pitfall:* Including/excluding outliers in analysis.

*Solution:* Verify outlier type, determine importance, check distribution without outlier, conduct analysis with and without. Include the outlier unless you have a reason not to.

*Pitfall:* Jumping straight to non-parametric methods and believing they have no assumptions.

*Solution:* Consider a transformation first and if not check assumptions (non-parametric tests have fewer assumptions).

Hypothesis Testing

*Pitfall:* Running multiple t-tests on the data.

*Solution:* Either use corrections or use a suitable model. If you run too many t-tests, you will eventually end up finding statistical significance. However, this difference may have shown up due to chance rather than a true scientific phenomenon.

*Misconception:* If the data is analyzed enough it will produce the “right answer.”

*Truth:* While technically correct, this is not what statistical testing should be used for. It should give the correct answer, but not necessarily the “right answer.”

Modeling

*Misconception:* A statistically significant difference is always important.

*Truth:* “Statistical significance” depends on the practical difference. If it influences a decision, it may be significant.

*Pitfall:* Analyzing summary statistics instead of raw data.

*Solution:* It is often better to see the data “as is” so the statistician and researcher can decide together whether something is missed by aggregating data or dropping observations, and all the data can be used for an interpretation.

*Pitfall:* Proceeding as if the data follow the experimental design when in reality data are missing or were taken under different conditions.

*Solution:* Start with the data “as is.” Consider carefully what will be lost by simplifying; your results will be more accurate. If you do not know how to analyze the data, consult a statistician.

Conclusions

- Plan your analysis.
- Document your steps.
- Consult a statistician before you begin your project, preferably at the project design stage.
- Always plan ahead.

These misconceptions and truths and pitfalls and solutions related to the analysis of data were adapted from Victoria Cox’s presentation, “Common Pitfalls and Misconceptions in Statistics with Suggested Solutions” from the Conference on Statistical Practice, San Diego, CA, February 23, 2016.
DATA STORAGE
Research Cyberinfrastructure’s key goals regarding data storage are: to provide long-term storage for research data for backup, preservation, and discovery; and to facilitate easy, secure, and fast sharing of publicly-funded research project results. University IT and Research Cyberinfrastructure currently offer three data storage solutions to MSU researchers.

**Box**
Unlimited storage for faculty, staff and students.
- Excellent for small datasets (less than 10GB); okay up to 500GB.
- Great integration with office tools.
- Automatic synchronization for Windows and Mac.
- Backed up and redundant.
- Not good when data comes from instruments or High Performance Computing (HPC).

**One Drive**
5TB storage for faculty, staff and students.
- Excellent for small datasets (less than 10GB); okay up to 500GB.
- Great integration with office tools.
- Automatic synchronization for Windows.
- Backed up and redundant.
- Not good when data comes from instruments or HPC.

**Hyalite Cluster**
600TB storage for Hyalite users.
- Excellent for computer-generated data.
- No integration with office tools.
- No automatic synchronization for Windows or Mac.
- Not backed up (but this will change soon).
- Not good for long-term storage nor sharing.
DATA PUBLICATION AND DISSEMINATION
Data Publication Requirements

Federal funding agencies, private funders, and publishers are increasingly requiring open publication of supporting research data alongside published articles.

Recommended data repositories

We recommend publishing data in a data repository in your discipline. If you have questions about where to publish data, please see our website (https://www.lib.montana.edu/services/data/publication) or contact the Data Librarian.

Recommended metadata fields when publishing data

We recommend providing the following descriptive elements for your data:

- Title of the dataset
- Creator (the person/people who collected the data)
- Date last modified
- Version, if applicable
- Short description of the dataset
- Citation and DOI for related publication(s), if applicable
- Licensing information. Copyright for research data is not always clear, since facts are not copyrightable. We therefore recommend that you commit your data to the public domain, using a Creative Commons Zero designation. For more information: https://creativecommons.org/about/cc0
  https://wiki.creativecommons.org/wiki/CC0_use_for_data

- We also recommend creating a codebook or readme—a plain text file that provides more information about your dataset.
- Codebook/Readme resources:
  - Checklist from Mozilla Science: http://mozillascience.github.io/checklist
  - Readme template, adapted from the University of Minnesota: https://goo.gl/9Lz5RB

Data Citation Practices

Citing a dataset is as important as citing a publication. If citing your own data in an associated article, refer to the dataset twice:

1. Include a data availability statement in the article text:
   Data available from Dryad Digital Repository, https://doi.org/10.5061/dryad.ab123

2. Include a full citation in the reference list:

When citing someone else’s dataset, cite the dataset as you would a publication: an in-text citation with author and year, and a full citation in the reference list. Include a DOI for the dataset if possible.
**Conclusion**

By managing data well, you can

- Make grant submissions viable.
- Benefit the public and scientific communities.
- Advance the frontiers of knowledge faster.
- Ensure protection of human subjects.

Another helpful guide: [http://data-archive.ac.uk/media/2894/managingsharing.pdf](http://data-archive.ac.uk/media/2894/managingsharing.pdf)
CONTACTS
**MSU Library Data Services**
Sara Mannheimer, Data Librarian  
406-994-3361 | [http://lib.montana.edu/services/data](http://lib.montana.edu/services/data)

Library Data Services help researchers create and implement a data management strategy in order to produce better research and comply with requirements from publishers and funding agencies. When research data is ready to be published, the library can help publish data in a data repository.

**Research Cyberinfrastructure**
406-994-3416 | [www.montana.edu/rci](http://www.montana.edu/rci)

Research Cyberinfrastructure (RCi) aims to provide MSU Researchers with the cyber-infrastructure and support they need to remain on the cutting edge of their fields. This is a collaborative effort between the UIT, the Library, and the Office of the Vice President for Research to provide additional research-specific resources that lie outside the scope of UIT services.

**Human Ecology Learning & Problem Solving (HELPS) Lab**
Eric Raile, Director  
406-994-4107 | [http://helpslab.montana.edu](http://helpslab.montana.edu)

The HELPS Lab is a fee-for-service facility at MSU-Bozeman that enables the collection of high-quality data for researchers employing a variety of social and behavioral methods. The HELPS Lab is open to the broader community of researchers, with an emphasis on providing tools necessary for researchers to study interactions between human systems and other complex phenomena like ecosystems and public health. The HELPS Lab facilitates, on a fee basis, the collection of high-quality data via computer-based experiments; interviews; focus groups; and web, mail, personal and phone surveys. The HELPS Lab also handles data entry, cleaning and documentation. Additional services, like transcription or assistance with sampling are also available on a fee basis.

**Statistical Consulting and Research Services**
Mark Greenwood, Director  
406-994-5594 | [www.montana.edu/statisticalconsulting](http://www.montana.edu/statisticalconsulting)

Statistical Consulting and Research Services (SCRS) offers statistical consulting to MSU students, faculty, as well as non-affiliates, dependent on consultant availability. SCRS is currently supported by Montana INBRE and Mountain West CTR-IN. SCRS can assist in the following areas:

- Research Planning: study design, sample size, statistical power, analysis planning, data collection, data management and research proposals.
- Analysis: computed variables, selecting statistical methods, modeling, model interpretation, and code review (R, SAS, or SPSS).
- Dissemination: graphical and tabular displays, model interpretation and writing.

**Office of Sponsored Programs**
406-994-2381 | [www.montana.edu/research/osp](http://www.montana.edu/research/osp)

The Office of Sponsored Programs (OSP) manages all functions related to sponsored programs at Montana State University defined as any activity, research or otherwise, that is funded by an external source such as federal, state, or private organization. The staff of OSP, comprised of 10 professional Fiscal Managers and a Pre-Award Specialist, takes a comprehensive managerial approach and engages in the award life cycle from the time a funding source is identified to the close out of a grant or award. The Proposal Services staff are available to work with PIs on planning, developing and submitting grant proposals, as well as providing referrals to people and programs who help with particular aspects of proposals.