NEBRASKA ACADEMY FOR
METHODOLOGY, ANALYTICS & PSYCHOMETRICS

# Ensuring Data Integrity Through Ongoing Maintenance and Monitoring

Jungwon Eum, Ph.D. & Amanda Prokasky, Ph.D

University of Nebraska Lincoln

# Outline of Talk

- Introduction to Data Integrity in the Social Sciences
  - Definition of Data Integrity
  - Importance of Data Integrity
- Implementing Data Integrity Throughout the Research Data Lifecycle
- 10 Best Practices to Ensure Data Integrity
  - Practical step-by-step guide of ongoing monitoring & maintenance

MAP ACADEMY

# Why Data Integrity Training?

- We need a systematic and ongoing method for ensuring data integrity
- There is no one standardized practice to assess data integrity
- Publications/presentations/final reports do not verify the data and how the data were processed
- There are limited training opportunities/resources/guidelines to ensure data integrity
- We need to intentionally plan and implement data integrity during a research project

# Definition of Data Integrity

- Domain/discipline specific:
  - Academia- focus is on management of scientific research data
  - Business world- focus is on security of data (e.g., trade secrets, forecasting)
  - Health fields- focus is on quality of data (e.g., FDA)
    - ALCOA
- Reliability and trustworthiness of data
- **Can I use the data vs can I trust the data**


Condon, P., Simpson, J., and Emanuel, M. (2022) Research data integrity: A cornerstone of rigorous and reproducible research, IASSIST Quarterly 46(3), pp. 1-21. DOI: https://doi.org/10.29173/iq1033

MAP ACADEMY

# Key Elements of Data Integrity

- **Data Management:** A set of foundational practices for organizing, documenting, storing, sharing, and preserving data

- **Data Quality:** 'Assurance that the data produced is exactly what was intended to be produced and fit for it's intended purpose' (Medicine & Healthcare Products Regulatory Agency, 2018, p/.20)

- **Data Security:** Physical security and technological protection of data for safeguarding data from corruption, unauthorized access, or loss

MAP ACADEMY

# When should we check data?

- Before data collection
- During data collection
- After data collection


Research Data Life Cycle
(https://researchdata.unl.edu)

# When should we check data?

- Before data collection (YES!)
  - Plan how to monitor data
    - Data monitoring plan
- During data collection (YES!)
  - Actively/Regularly monitor data
    - Data monitoring checklist/protocols
- After data collection (YES!)
  - Finalize data for analysis
    - Analysis-ready datasets


Research Data Life Cycle
(https://researchdata.unl.edu)

# Unified Definition of Data Integrity

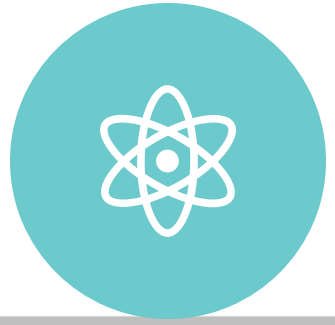
- The degree to which data are complete, consistent, accurate, trustworthy, reliable and that these characteristics are maintained ***throughout the data life cycle***


MAP ACADEMY

# Importance of Data Integrity: The Four Rs of Research

**Rigor-** strict application of the scientific method to ensure unbiased and well-controlled experimental design, methodology, analysis, interpretation and reporting of results (US Department of Health and Human Services)

**Reproducibility-** consistent computational results using the same input data, steps, methods, code, and conditions of analysis (National Academies of Sciences, Engineering, and Medicine, 2019)



**Replication-** consistent results across studies aimed at answer the same scientific question with their own data (National Academies of Sciences, Engineering, and Medicine, 2019)

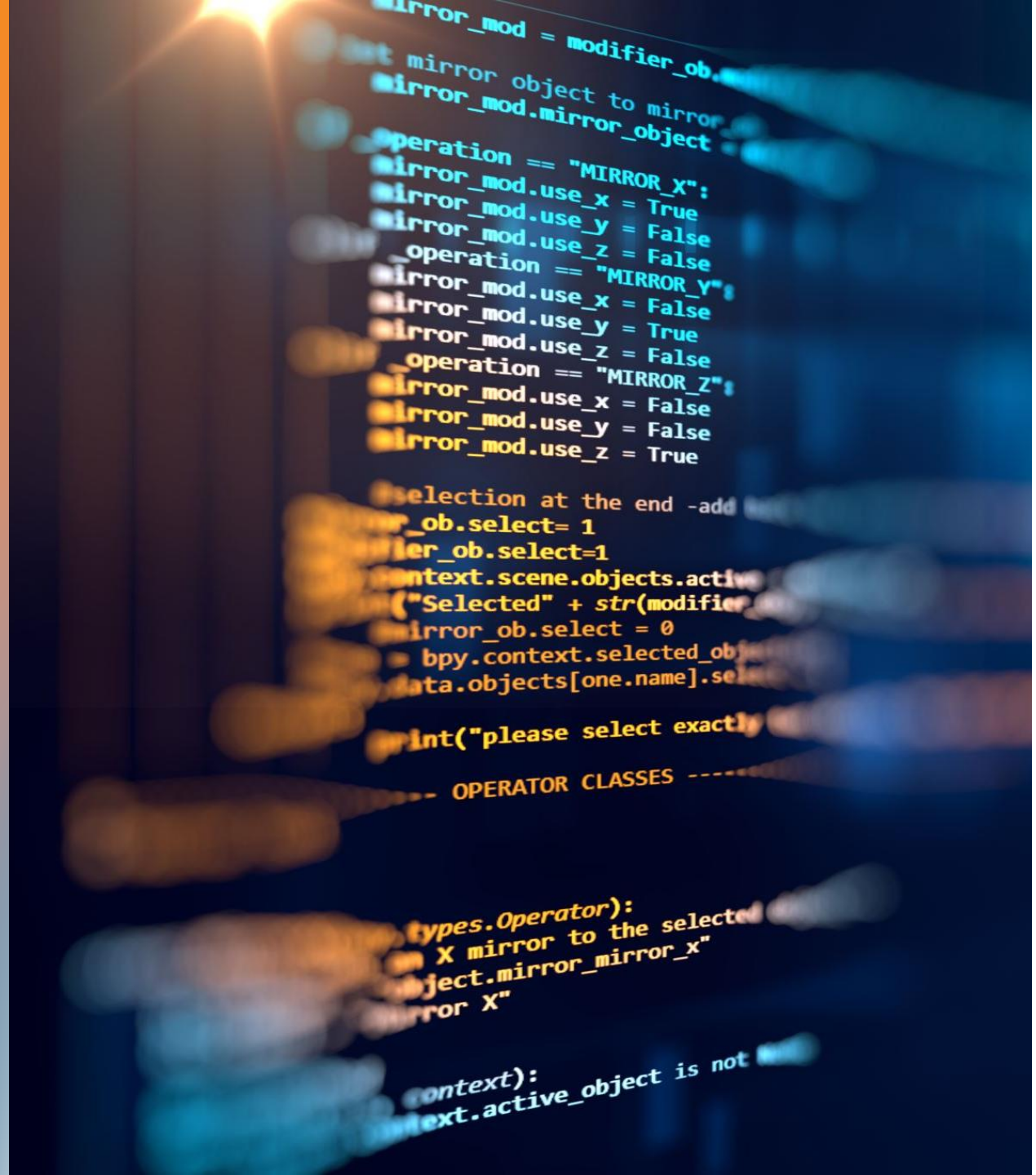
**Reuse-** use research data for a research activity or purpose other than that for which it was originally intended (National Library of Medicine)


# Importance of Data Integrity

- Valid research findings
- Policy-making
- Ethical responsibility
- Accurate reporting
  - CONSORT Chart

# Consequences of Poor Data Integrity

- Inaccurate results, flawed conclusions
- Contributes to replication crisis in social science research
- Loss of public trust
- Wasted resources and funding



# Common Threats/Challenges to Data Integrity

- Data Collection Challenges
  - participant non-response, measurement errors
- Data Entry and Management Errors
  - human error: entry, coding, initial processing
- Longitudinal studies
  - attrition, consistency in measurement

# Principles of Ongoing Maintenance for Data Integrity

Data Cleaning and Validation- regular procedures for identifying and correcting inaccuracies

Documentation Standards- importance of detailed documentation of data collection methods, coding decisions, and transformation processes

Version Control- strategies for maintaining versions of datasets, especially in collaborative research

# Essential Elements of Data Monitoring Plan

Who is responsible for performing the data monitoring?

How often will monitoring be performed?

What components of the study will be monitored?

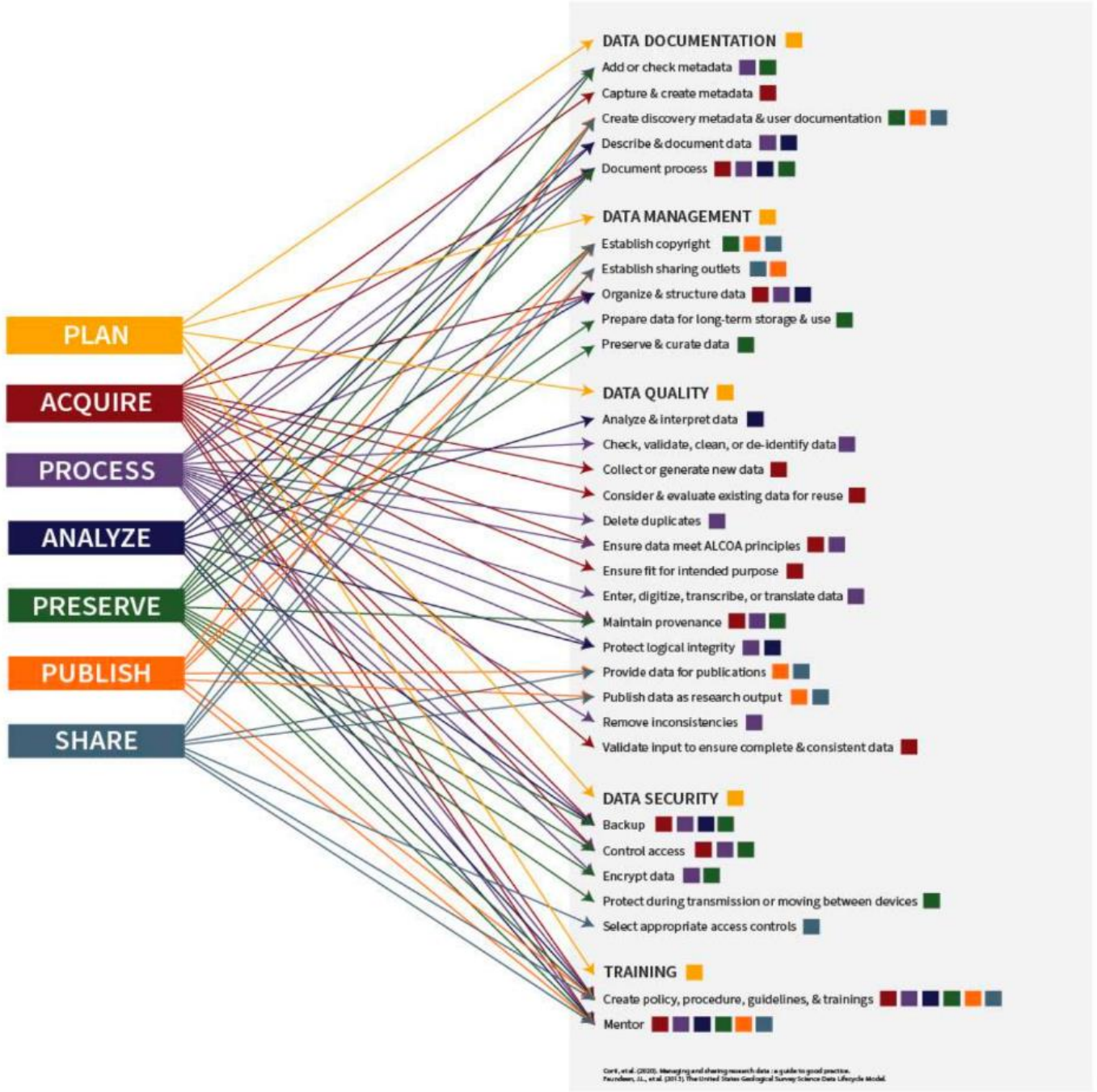
How will monitoring be documented and responded to appropriately?

Before

After

# Data Integrity Implementation Schema

Condon et al (2022)



# Implementing Data Integrity throughout the Research Data Lifecycle



Each stage of the Research Data Lifecycle is associated with activities that correspond with multiple of the components related to data integrity.

e.g., data quality is not achieved at a single point in the Research Data Lifecycle, but rather through multiple activities that take place throughout the lifecycle.



Researcher actions and the data integrity components are not isolated activities, but rather they interact with one another to build a web for achieving data integrity.

# Implementing Data Integrity throughout the Research Data Lifecycle

It is significant to note that the Plan stage uniquely maps directly to each of the key components and not to specific actions.

Researchers should plan for the acquisition, processing, analysis, publication, sharing, and preservation of data.

Planning for **data management**, **data quality**, **data security**, as well as **documentation** and **training** needs to occur.

# Implementing Data Integrity throughout the Research Data Lifecycle

Purposefully planning for research data integrity requires careful consideration of its components and their relationships to yield reliable, trustworthy, valid, and secure data throughout each stage of a Research Data Lifecycle.

Purposeful planning enables researchers to conduct rigorous research and generate outcomes that not only reflect the desired research data integrity characteristics, but that are also reproducible, replicable, and reusable.

# Implementing Data Integrity throughout the Research Data Lifecycle

What is also noticeable in the Plan stage, along with Acquire and Process stages, is that most of the activities associated with **data quality** fall into these three stages of the lifecycle.

This is because researchers are actively working with the data.

e.g., collecting/generating, cleaning, manipulating, and interpreting data

At the Preserve stage the quality of the data has been established so, addressing integrity focuses on maintaining that quality through security and curation (long-term data management) of the data to prevent unauthorized changes.

# Implementing Data Integrity throughout the Research Data Lifecycle

Mapping between the Research Data Lifecycle stages and activities associated with research data integrity core components illustrates the **complexity of research data integrity in practice**.

There is not one action that researchers take to achieve research data integrity or, for that matter, one action to achieve quality, security, or the management of data.

# 10 Best Practices to Ensure Data Integrity


MAP ACADEMY

# Practice 1: Implement Robust Data Validation Processes



**Regularly check data for accuracy, completeness, and consistency (CONTINUOUS MONITORING is needed! – e.g., monthly)**

Is everything making sense?; Data are accurate, complete, and consistent?;Variables are coded consistently?; Variables are formatted correctly?



**Check data consistency internally and against master file (e.g., codebook)**

After cross-checking, further investigation (digging) or error reports for invalid, questionable, or inconsistent data may be needed.

Data Quality

MAP ACADEMY

# Practice 1: Implement Robust Data Validation Processes

**Check data as they come in!**

**Don't wait until you finish data collection!!**

**Check the following:**

| | | | | | | |
|---|---|---|---|---|---|---|
| Measures | Variable names | Variable responses /values<br>• Numeric? Text? | Variable labels<br>• Need key info (e.g., teacher reported PTRS 1. xxx at time 1) | Response value labels<br>• e.g., 0 = No; 1 = Yes or Likert scale 1 = strongly disagree; 4 = strongly agree | Missingness patterns<br>• Consistent? Inconsistent? Makes sense? | Missing values and labels<br>(e.g., -999 = non-response; -888 = don't know; -777 = refusal of item) |

Data Quality

MAP ACADEMY

# Practice 1: Implement Robust Data Validation Processes

Range and validity checks (check any data entry errors, data anomalies; run frequencies/descriptives)

- Check for impossible or extreme scores (through descriptives/frequency analyses or data visualization)
- e.g., Values should be ranged from 1 to 5, but there is an impossible value (e.g., 6).
- e.g., Mean range should be1-5 but the mean score was 7.56.

Communicate data issues with the team as soon as possible (any anomalies, inconsistences, problems)

Document all the decisions made, changes with details

- Who, when, where, what, why, how

Data Quality

MAP ACADEMY

# Practice 2: Audit



Auditing is a critical aspect of data integrity assurance.



Auditors must have specific knowledge, insights, and skills.

Data Quality

MAP ACADEMY

# Practice 3: Implement Good Data Management Practices



Data management entails entering, cleaning, scoring, and processing data.



High quality data management practices focus on **reducing the amount of error** added at any of the data life cycle stages.



Double checking, cross checking, documenting, using syntax (for tracking & replicating), being thorough

Data Management

MAP ACADEMY

# Practice 3: Implement Good Data Management Practices



## Data Entry

Independent data entry and checks against the original forms when discrepancies are detected



## Data Cleaning

De-identification, range checks for items to ensure values for all items and scale scores are valid



## Data Scoring

Computer scoring of all measures (sum scores, mean scores, T-scores, Standard scores)



## Data Processing

Consistency checks within and across datasets, creation of new versions of datasets as identified errors in the data are corrected, merging by data level (e.g., teacher level, child/family level)

Data Management

MAP ACADEMY

# Recommended Standard Data Management Practices



**Step 1: Export data**



**Step 2: Preserve original data (e.g., Originals – DO NOT TOUCH)**

Original data should never be worked directly



**Step 3: Make copies of the original files and work with the copies**



**Step 4: De-identify, clean, and check data (through cleaning syntax)**

An error report is issued whenever discrepancies are found.

This report becomes part of the documentation for the project, all the changes and decisions made during data processing.

Further investigation (digging) may be needed.

Create one master file that includes all the detailed information (corrections, changes, decisions made, what changes were made and when, whether a change was made)



**Step 5: Score data (through scoring syntax)**

Scoring instructions are needed (e.g., codebook)

Run checks on the items to make sure that the data was collected correctly



**Step 6: Save data**

Data Management

MAP ACADEMY

# Common Examples of Data Cleaning Activities (when applicable)

- Remove unnecessary, identifiable, confidential variables
- Remove cases that were part of testing, duplicate entries, cases without usable data, cases without consent
- Make sure missing data is coded appropriately
- Check for impossible/improbable values
  - e.g., a question that asks about the age and there are answers >100
- Check for and make decisions about outliers
- Make sure text entry data is formatted correctly
  - e.g., changing ages written as words into integers (twelve -> 12)
- Create new versions of variables to meet needs

Data Management

Source: https://researchdata.unl.edu

MAP ACADEMY

# Common Examples of Data Cleaning Activities (when applicable)

- Make sure all numerical values use the same metrics
  - e.g., data on time, weight, length, currency, temperature, etc. may all be entered into different types of units
- Check the coding of partially closed ended responses
  - e.g., other, please specify
- Make decisions about bad faith answers
  - e.g., responses that do not appear to be a sincere attempt to answer the question asked
- Make decisions about whether to remove cases due to poor data quality
- Look for evidence of bots (online data collection)

Source: https://researchdata.unl.edu

# Practice 4: Use Access Control



**Limit the number of people who can access or perform actions on data to reduce the risk of human error**



**Limit both physical and online accesses**

# Practice 5: Back up Data



## Regularly take a backup of your data



## 3-2-1 rule

3 copies of the data (e.g., original+ external/local + external/remote)

2 types of storage formats (e.g., external hard drive + cloud location)

1 storage type is offsite (e.g., geographically distributed location)



MAP ACADEMY

# Practice 6: Keep an Audit Trail



**Whenever there is a problem, it's critical to be able to track down the source.**



**An audit trail provides the team to accurately pin point the source of the problem.**



**Every event – create, delete, read, modified, when, by whom – is tracked and recorded automatically.**

e.g., OneDrive allows to keep track of every event and restore an old file if needed



MAP ACADEMY

# Practice 7: Train Staff on Data Handling Procedures



**Ensure employees understand the importance of data integrity and follow proper protocols**

e.g., how to manage the data, how to preserve the data quality, how to protect confidentiality



**Staff should have the ability to understand, handle, manage, and ethically use data.**



MAP ACADEMY

# Practice 8: Establish Collaboration

Define roles and responsibilities

Keep all the team members of your project on the same page

Everyone should be in the loop.

It will allow you much-needed control over your data.

When collaborators are operating with the same philosophy and vocabulary about data, a shared understanding of each collaborator's responsibilities, activities, and outputs is more easily achieved.

# Practice 9: Document all

**Document all the information necessary to understand the content and context of the data**

**Document all decisions and changes made (e.g., recoding variables, correcting errors, scoring decisions)**

- Update documentation (e.g., codebook) as changes occur
- Living document

**Project-level documentation**

- Data collection context & methodology, file structure and organization, data validation, quality assurance, etc.

**Data-level documentation**

- Variable names & descriptions, definition of codes, codes for missing values, processes used to clean data, etc.

Data Documentation

MAP ACADEMY

# Practice 10: Create a Tracking System



**Create a tracking system using protocols/syntax files/tables for data entry and data processing**



**Each step in entering, scoring, and analyzing the data should be completely documented so that it is possible to identify what was done in each stage.**

Data Documentation

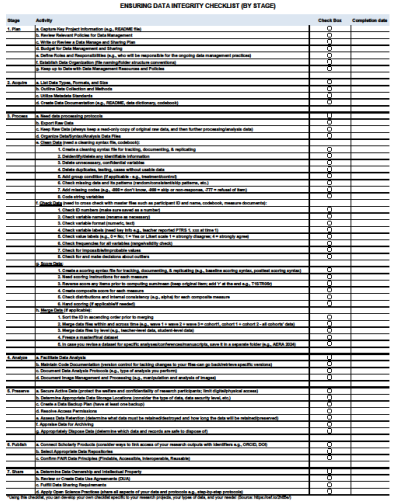
MAP ACADEMY

# Resources:

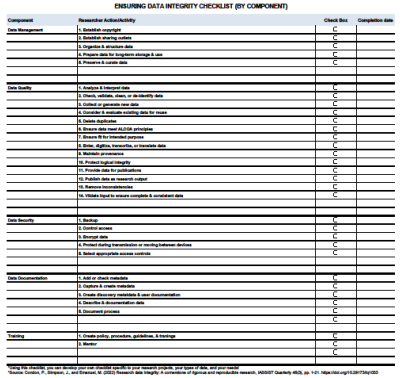
Scan me for more resources

Handouts

- Data Integrity Checklist

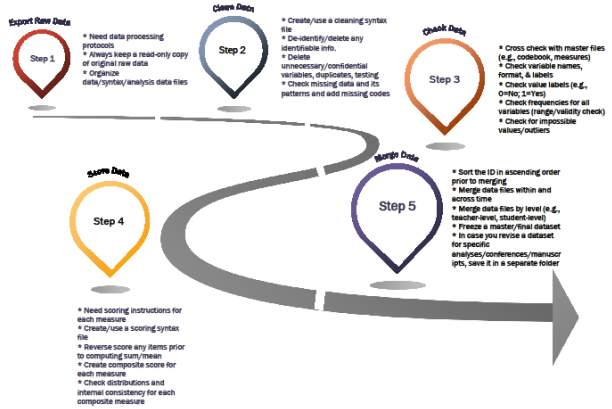


ENSURING DATA INTEGRITY CHECKLIST (BY STAGE)



ENSURING DATA INTEGRITY CHECKLIST (BY COMPONENT)

- Data Cleaning Steps




https://go.unl.edu/map_data_integrity
https://datamanagement.hms.harvard.edu/plan-design/biomedical-data-lifecycle

MAP ACADEMY

# References


- Burchinal, M., & Neebe, E. (2006). Best practices in quantitative methods for developmentalists: I. Data management: Recommended practices. *Monographs of the Society for Research in Child Development, 71(3)*, 9–23. https://doi.org/10.1111/j.1540-5834.2006.00354.x
- Condon, P., Simpson, J., and Emanuel, M. (2022) Research data integrity: A cornerstone of rigorous and reproducible research, *IASSIST Quarterly* 46(3), pp. 1-21. https://doi.org/10.29173/iq1033
- Corti, L., Van den Eyden, V., Bishop, L., and Woollard, M. (2020) Managing and Sharing Research Data: A Guide to Good Practice. 2nd edn. London: Sage Publications.
- McDowall, R.D. (2018) 'How to Use This Book and an Introduction to Data Integrity', in Data Integrity and Data Governance: Practical Implementation in Regulated Laboratories, pp. 1-27. Available at: https://doi.org/10.1039/9781788013277-00001
- Ng, C. (2021) 'What is Data Integrity and How Can You Maintain it?', Inside Out Security. Available at: https://www.varonis.com/blog/data-integrity (Accessed: 17 March 2022).
- University of Nebraska-Lincoln – Research Data. Available at: https://researchdata.unl.edu
- University of Nebraska-Lincoln – Libraries. Available at: https://libraries.unl.edu/research-data-management/
- Harvard Longwood Medical Area Research Data Management Working Group: Project Work. available at: https://osf.io/2h65e/


MAP ACADEMY

NEBRASKA ACADEMY FOR
METHODOLOGY, ANALYTICS & PSYCHOMETRICS



THANK YOU!



QUESTIONS?



CONTACT: JEUM@UNL.EDU,
APROKASKY3@UNL.EDU