

#### NEBRASKA ACADEMY FOR METHODOLOGY, ANALYTICS & PSYCHOMETRICS

# Ensuring Data Integrity Through Ongoing Maintenance and Monitoring

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



## **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



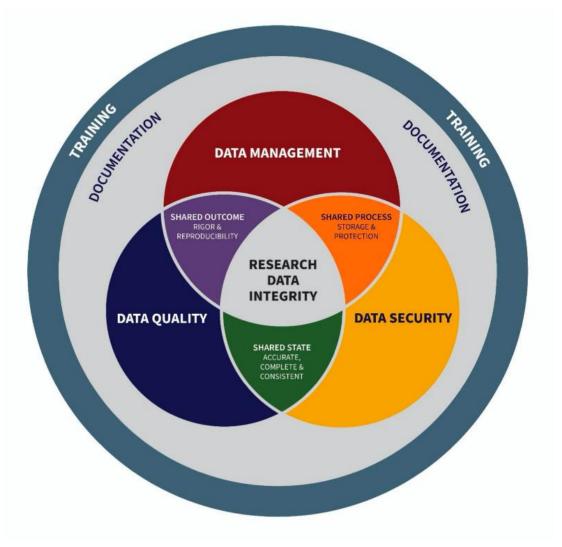
#### Why Data Integrity Training?

- 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



## **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



## 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

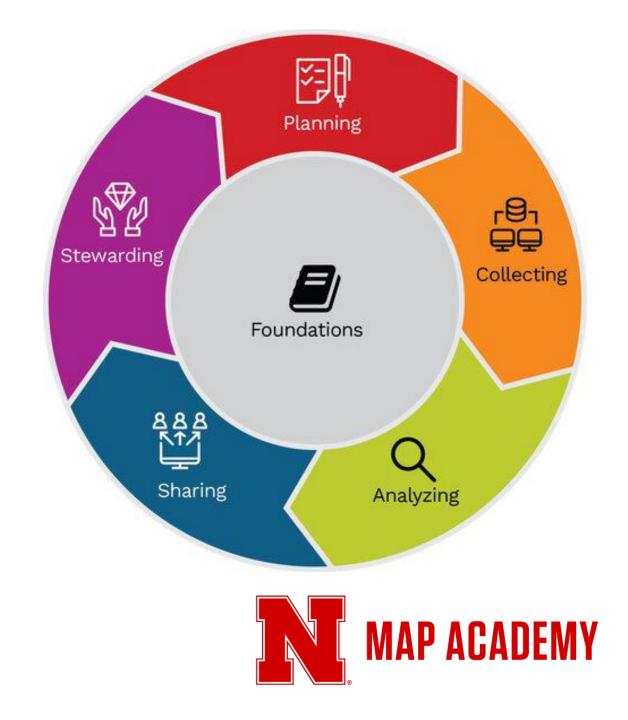


Research Data Life Cycle (https://researchdata.unl.edu)



#### Unified Definition of Data Integrity

The degree to which data are complete, consistent, accurate, trustworthy, reliable and that these characteristics are maintained <u>throughout the</u> <u>data life cycle</u>



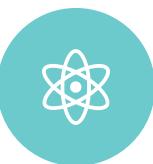
# **Importance of Data Integrity: The Four Rs of Research**



**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)

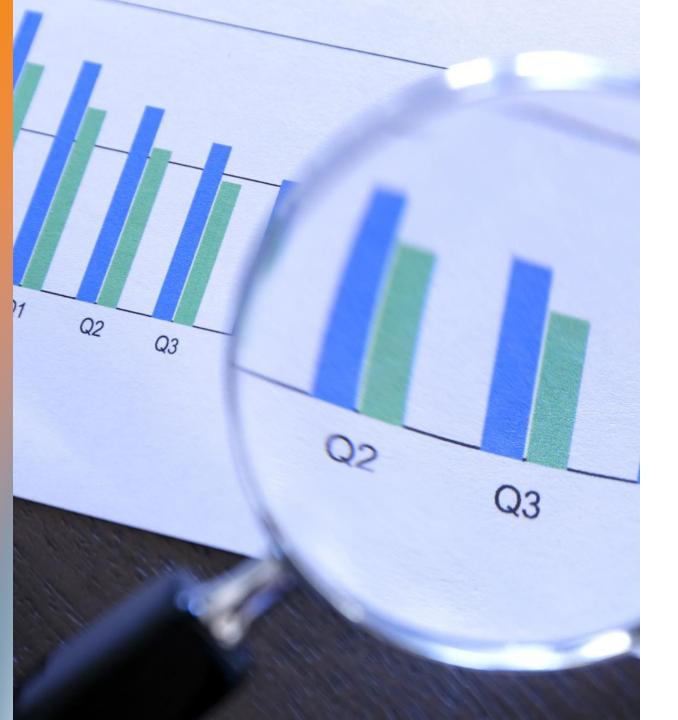


**Replication-** consistent results across studies aimed at answer the same scientific question with their own data (National Academies of Sciences, Engineering, and Medicine, 2019)



**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**





election at the end -add \_ob.select= 1 er\_ob.select=1 ntext.scene.objects.action "Selected" + str(modifient irror\_ob.select = 0 bpy.context.selected\_ob ata.objects[one.name].selected\_ob ata.objects[one.name].selected\_ob

plnt("please select exactle

-----

ypes.Operator): X mirror to the select ject.mirror\_mirror\_x" ror X"

ontext):
 object is not
 active\_object is not

#### **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?



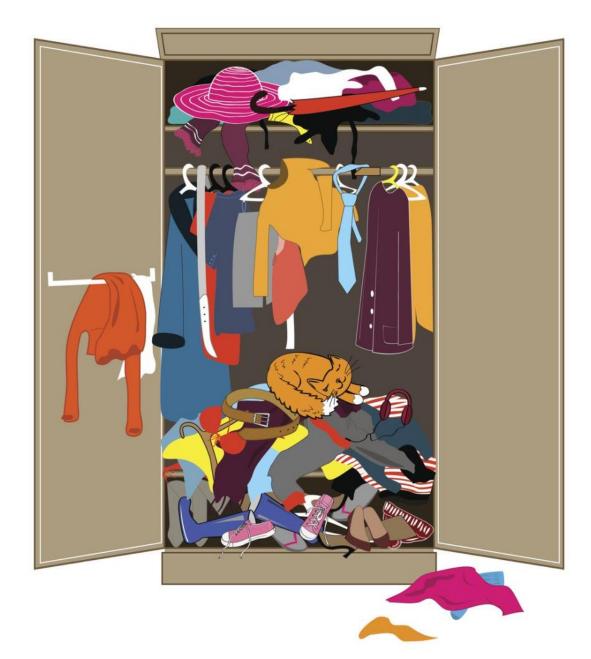
What components of the study will be monitored?

How will monitoring be documented and responded to appropriately?









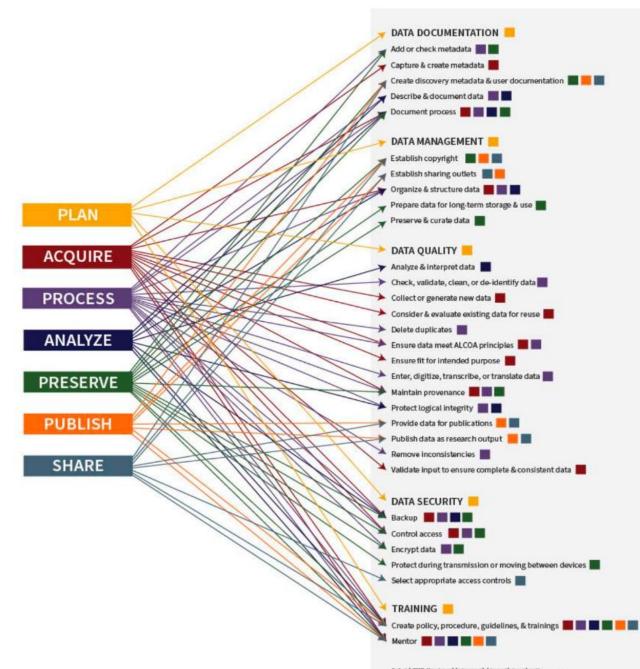






#### Data Integrity Implementation Schema

Condon et al (2022)



Cort, et al. (2020). Menaging and sharing mounth data : a guide to good practice. Foundeer, JL, et al. (2013) The United Status Geological Survey Science Data Liberycle Model.



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.

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.



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.



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



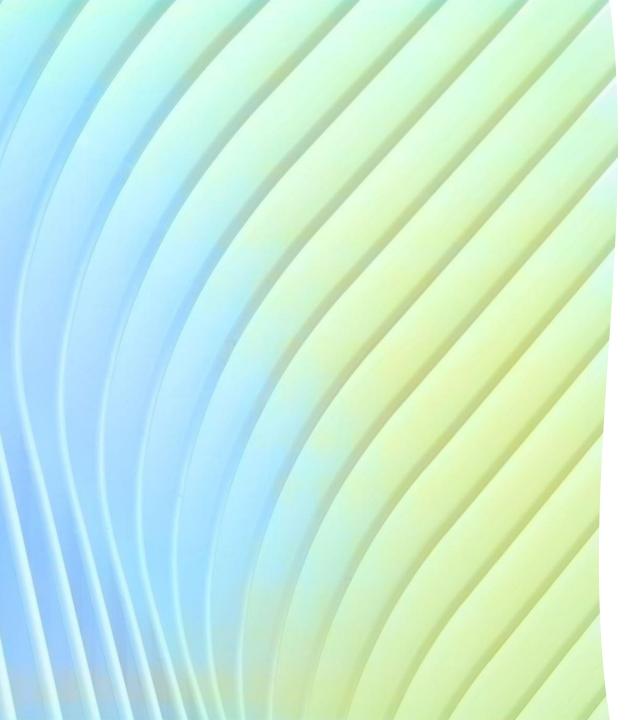
At the Preserve stage the quality of the data has been established so, addressing integrity focuses on maintaining that quality through security and curation (long-term data management) of the data to prevent unauthorized changes.



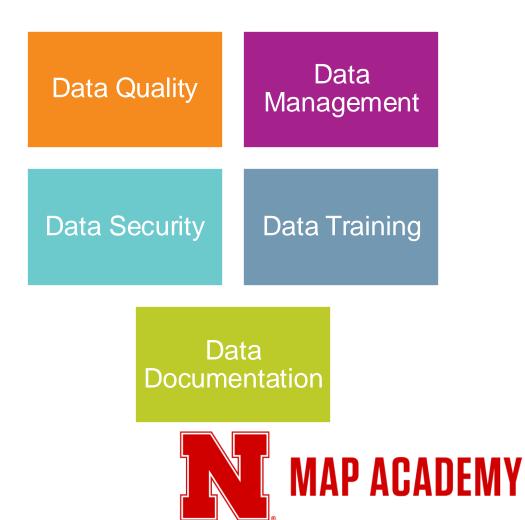
Mapping between the Research Data Lifecycle stages and activities associated with research data integrity core components illustrates the **complexity of research data integrity in practice**.



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**



### 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?

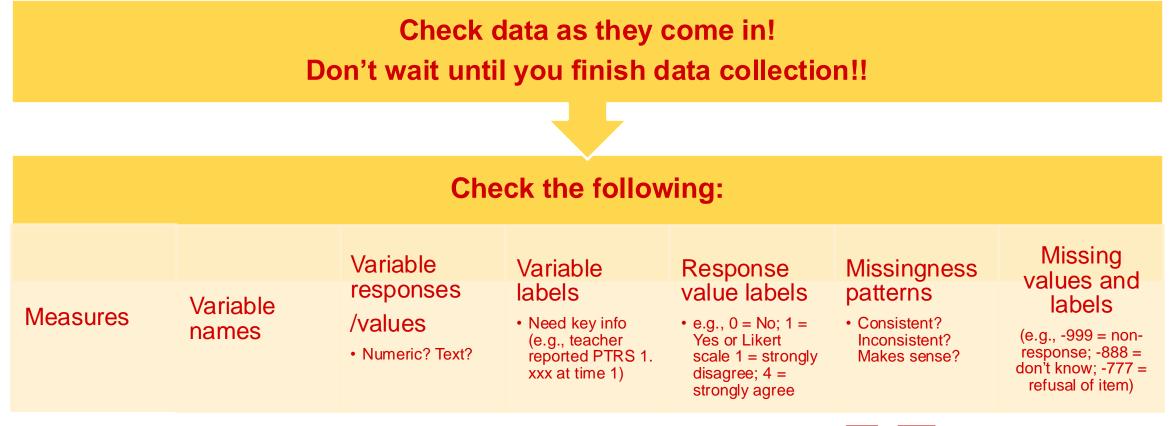


#### Check data consistency internally and against master file (e.g., codebook)

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



### Practice 1: Implement Robust Data Validation Processes





### Practice 1: Implement Robust Data Validation Processes

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

- 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



**Practice 2: Audit** 



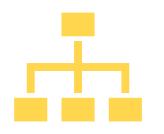


Auditing is a critical aspect of data integrity assurance.

Auditors must have specific knowledge, insights, and skills.



# **Practice 3: Implement Good Data Management Practices**







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

Data

Management

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



# **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

### **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

1

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

**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

Data

Management

- 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



## **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

Data

Management

- 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

**MAP ACADEMY** 

#### **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



Data Security

### **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)



Data Security

#### **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



Data Security

### **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.



Data Training

### **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.



Data Training

### **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

### **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

#### **Resources:**

#### Handouts

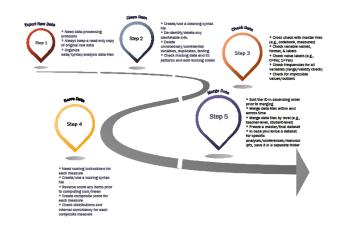
- Data Integrity Checklist

			Consistion date
Page 1	Automy	Check Bas	Comptetion data
100	a Captum Key Project Information (a.g., READART file) D. Review Relevant Policies for Data Management	- X	
	c. Write or Review a Data Manage and Sharing Plan	- X	
	4. Budget for Dela Management and Sharing	<u> </u>	
		0	
		0	
	p Famp up to Date with Date Management Resources and Publics		
L Acquire	a. List Data Types, Formate, and Disa	0	
	h Cutine Dels Collection and Methods	0	
	n Villes Meindels Garciaris	0	
	6 Create Data Documentation (e.g., NEACHE, data distionary, codebook)	0	
Process	A final data processing protocola		_
PRODUCT	a Nade data producing protects b. Export Ray Data	8	_
	p. Export New Cells (always keep a med-only copy of original raw data, and then further processing/analysis data)	- X	
		ŏ	
	<ol> <li>Common Deterfysicationarysis Data Pleas</li> <li>Chem Data (need a chaming syntax file, codeboot):</li> </ol>		
	Constex cleaning system the for backing, documenting, & replicating	8	
	<ol> <li>Debite any second s</li></ol>	8	
_	4. Delate duplicates, teating, cases without cashin data	0	
_		0	
_	6. Check missing data and its patterns (random/sonaistentikilo patterns, etc.)	0	
	7. Add minning codes (s.g., 400 = don't know, 400 = skip or non-maponan, JTT = minaal of karr)	0	
	E Code atring variables	0	
	<ol> <li>Obsci: Data lived to once check with master files such as participant ID and name, codebook, measure documents):</li> </ol>		
	1. Check ID numbers (make sure served as a number)	8	
	Oherk variable names (entens as nacessary)     Oracle variable formal (in made, both	- X	
		8	
	<ol> <li>Check variable labels (need key into e.g., beacher reported PTRS 1, xxx at time 1)</li> <li>Check value labels (e.g., 0 = No; 1 = Yes or Libert scale 1 = attorphy disagree; 4 = attorphy agree)</li> </ol>	8	_
	6. Check Inspandes for all raciables (angehold by check)	ŏ	
	7. Check for impossible-improbable values	8	
	E Charle for and make decisions about output	8	
	s Jaan Dela		
	<ol> <li>Create a scoring syntax file for building, documenting, &amp; replicating (e.g., baseline acceling syntax, positive) acceling syntax)</li> </ol>	0	
	2. Nand souring instructions for each managem	0	
	3. Revense access any litena prior is computing sumitmen (keep original item; add Y at the and e.g., 7457805)	0	
	4. Create composite score for each measure	0	
	5. Check distributions and internal consistency (e.g., sibility for each composite measure	8	
	6. Hard scoring (Fapplicable) maded) Is Herps fairs (Fapplicable)		_
	1 Sort the ID in according order prior to merging	0	
	Here data fire within and across firms (e.g., www.1.+ www.2.+ www.3.+ cohort1, cohort1.+ cohort2 all cohorts' data)	ŏ	
_	Merge data film by level (e.g., techer-level data, student-level data)	<u> </u>	
	4. Freeze a manifer final datament	ŏ	
	5. In case you revise a dataset for specific analyses/conferences/manuac/gits, save 1 in a separate folder (e.g., ASRA 2006)	8	
Analyse	A. Facilitate Data Analysis	0	
_	5. Mehdels Code Deconvertation (version control for tacking changes to your files can go backheb leve specific ventors)	0	
	n Decumenti Dala Analysia Protocola (a.g., type of analysia you parform)	0	
	4 Document Image Management and Proceeding (e.g., manipulation and analysis of images)	0	
	<ul> <li>Secure Active Cale (protect the wallers and confidentiality of research participants; limit digitality/usion access)</li> </ul>		
Preserve		8	
	b. Determine Appropriate Data Discrept Locations (monifor the type of data, data security level, etc.)	8	
	je, Crielle a Dina Bactup Pari (save in wait one bactup) 14 Baceba Acreas Parminsiona	8	
	p. Recover Access Permissions is Assess Data Relation (determine what data must be related destroyed and how long the data will be related preserved)	8	
	je Asses (als Falenciar (searchine and cals must be rearred betroyed and for long the data will be rearred preserved) If Approve Oats for Anthrony	- X	
	a Appropriately Dispose Data (determine which data and records are safe to dispose of)	8	-
			-
D000	a Connect Scholarly Products (consider ways to finit access of your research outputs with Identifiers e.g., OROID, DOI)	0	-
-	D. Select Appropriate Data Repositories	8	
	(. Confirm FAIR Data Principles (Findulis, Accessible, Interopendia, Resultin)	8	
Chara	a Deleventre Dela Cenanship and Intellectual Property	i u	1
	h. Ranker er Creale Data Use Agneeneris (2014) n Fulfil Data Sharing Regularmente	0	

Component	Researcher Action/Activity	Check Box	Completion date
Component Data Management	1. Examine any lot		Companyor and
Lora Management	1. Establish sharing cullate	2	
	5. States in Starvig Colors		
	Organize & advantase data     4. Prepare data for long-term atompe & une		
	E. Preserve & carole data		
Carla Quality	5. Analyze & Interpret data		
	1. Charly validate, staar, or de identify date		
	Cuthod: or generate new data     A. Consider & evaluate existing data for rease		
	<ol> <li>Consider &amp; evaluate schelling data für reute</li> <li>Datate Andronese</li> </ol>		
	5. Delete deploates 5. Delete deploates 5. Ensure dete meet AU/DA principles	с Г	
	7. Ename filts meet Autom principes		
	5. Eviter, digitize, transcribe, or translate data 3. Maldala provenance		
	10. Protect logical integrity		
	15. Provide data for publications	e e	
	12. Publish data as reasanth output		
	13. Remove inconsistencies	С	
	14. Vilidate legad to annuare complete & scralatent data	с	
Ovia Security	1. Bedup	C	
	2. Control access	C	
	A Energy data	C	
	4. Protect during transmission or moving batwase devices		
	5. Selecil appropriate access controls	c	
Data Documentation	1. Add or check metadata	c	
	2. Capture & create metadata	с	
	3. Ornate discovery metablete & user documentation	C	L
	4. Describe & decorrectistion dels	c	
	5. Document process	C	
		с	
Thinks	1. Ornate policy, procedure, guidelines, & tranings	С	
	2. Wanter	c	
		- C	



#### - Data Cleaning Steps





https://go.unl.edu/map\_data\_integrity https://datamanagement.hms.harvard.edu/plan-design/biomedical-data-lifecycle

### 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. <u>https://doi.org/10.1111/j.1540-5834.2006.00354.x</u>
- 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. <u>https://doi.org/10.29173/iq1033</u>
- 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: <a href="https://researchdata.unl.edu">https://researchdata.unl.edu</a>
- University of Nebraska-Lincoln Libraries. Available at: <a href="https://libraries.unl.edu/research-data-management/">https://libraries.unl.edu/research-data-management/</a>
- Harvard Longwood Medical Area Research Data Management Working Group: Project Work. available
   at: <a href="https://osf.io/2h65e/">https://osf.io/2h65e/</a>





