



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

# **Developing a Strong Research Data Infrastructure**

Lorey Wheeler, PhD

Methodology Applications Series

October 4, 2024



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

---

## Agenda

- Introduction
- Documentation
- Data Collection & File Creation
- Data Processing & Cleaning
- Data Storage & Access



# What is research data infrastructure?

- A system of tools, platforms, and processes designed to effectively collect, store, manage, and share research data across different stages of the research lifecycle
- Allows researchers to access and analyze data efficiently while ensuring its long-term accessibility and integrity

## Research Data Life Cycle



<https://researchdata.unl.edu/>

# What is data management?

- **Data management** includes acquiring, validating, storing, protecting, and processing required data to ensure the accessibility, reliability, and timeliness of the data for its users
- **Data management** entails everything from designing data collection tools, tracking, entering, processing, cleaning, scoring, and documenting data



DATA



KNOWLEDGE



ACTION



**Why is data  
management  
so important?**



# 1. Ensuring Integrity of Research Data

- One purpose of data management is to ensure quality of research data
  - Reduction of error in data
  - Retaining power to test hypotheses
- Non-optimal data management practices can introduce substantial amounts of error into the data
  - Often projects have strict training criteria on data collection techniques but no training in data management tasks
  - Example – can result in high rates of error
    - 55% of test scores incorrect
      - Entering wrong dates
      - Computing number of items passed incorrectly
      - Incorrectly entering ID numbers
      - Incorrect entry of scores into database



## 2. Open Science Requirements

- Standards are changing for how we conduct research...
  - Sharing data once a study is complete or in the publication cycle
    - APA journals: authors contract that data will be made available to peers for the purpose of verifying the findings
    - Publisher of Journal of Youth and Adolescence has instituted data sharing policies (Levesque, 2017)
    - Funding agencies (e.g., Institute of Education Sciences) requiring data sharing plans
- American Psychological Association (APA) partnering with Center for Open Science
- See Transparency and Openness Promotion Guidelines (TOP) (Nosek et al., 2015)



# 3. Federal Funding Requirements

- Data Management and Sharing (DMS) Plans
  - National Institute of Health
    - <https://sharing.nih.gov/data-management-and-sharing-policy>
  - Institute of Education Sciences
    - [https://ies.ed.gov/funding/datasharing\\_implementation.asp](https://ies.ed.gov/funding/datasharing_implementation.asp)
  - National Science Foundation:
    - <https://new.nsf.gov/funding/data-management-plan>
- Under the policy, investigators and institutions must plan and budget for the managing and sharing of data, submit a DMS plan for review when applying for funding, and implement the approved DMS plan.

# 4. Addressing Barriers to Sharing

- Barriers to Sharing Data...Room for Improvement
  - Limited and Ineffective Sharing
    - Only 27% shared some of their data – the remainder failed to comply
    - Of those who shared, only 48% had understandable data sets
  - Addressing Scholar's Concerns
    - Protection of participants, privacy
    - Problematic use of data by public

# Essential Components of Strong Data Infrastructure

- Thorough documentation starting in the planning phase
  - Project codebook and metadata
  - Syntax (code) as documentation
- Data collection and file creation protocols
  - Database design
  - Centralized data storage location
  - Tracking system
- Data processing and cleaning protocols
  - Preliminary analyses
  - Missing data protocols
  - Data de-identification (how and when to de-identify each data type)
- Data storage and access protocols
  - Centralized data storage location

(Burchinal & Neebe, 2006)



# Documentation



# What is documentation?

- **Documentation**: describes each step of the research process – from proposal to completion
  - Creation of comprehensive documentation is essential!
  - Facilitates dissemination and replication
  - Critical for supporting rigor, transparency, and open science
- **START DOCUMENTING AT THE INCEPTION OF PROJECT!**
  - This will save much time and many headaches
  - Prevents you from having to figure out later what happened

# Documentation takes many forms

- Study design and methods, instruments (measures), software
  - Protocols, scripts, recruitment materials, permission letters, informed consent forms
  - Original surveys/instruments
  - Codebook files
  - IRB protocols
- Syntax (code) files used to convert, clean, merge, score, and analyze data
  - Documents all data management and analytic decisions
- Labels used in data files
- Meetings notes, especially ones that include rationale for why specific decisions were made
  - E.g., system of OneNote Notebooks, sections, and pages
- Investigators/authors, affiliations, funding
  - Grant proposal

# Metadata

- Data about data
- Required for effective data sharing and use of data repositories
- Documentation:
  - Title, topic, and description of study
  - Study design and methods, instruments (measures), software
    - Sample and sampling procedures
    - Weighting and weighting procedures
    - Unit(s) of analysis/observation
  - Investigators, authors, affiliations, funding
  - Persistent Identifiers (ORCID, DOI, etc.)
- Describes:
  - Dataset as a whole
  - Data files
  - Data itself (observations)

# Variable Level Metadata

- Description of variables, expected values & decimal places, measurement units, how missing values encoded, constructed variables
- Who was asked and was not asked each question (skip patterns)
- Codebook, data dictionary, data catalog (may be built into software)
  - Here called a ‘codebook’
  - A document used to describe all variables collected in a study



# 1a. Codebook Contents

- Measure description and sources (citations), including demographics
  - Adaptations from original published measure
  - Changes in items over time and across reporters
  - Variable names and labels
  - Item text (all languages/translations)
  - Response choices and values (codes)
  - Missing data codes
  - Skip patterns (routing of questions)
- Participant instructions / introductory statements

# 1b. Codebook Contents

- Reporters (who reported on which measures)
  - Parents, youth, teachers
- Data (variable) sources
  - Survey data, observational data, physiological data
- Aspects of data collection (study design)
  - Timing / measurement occasions: waves
    - Wave 1 = Pre-intervention; Wave 2 = post-intervention
  - Phases
    - Phase 1 = Open pilot trial; Phase 2 = randomized controlled trial
  - Cohorts (groups of participants recruited at the same time)
    - Cohort 1 = Fall 2024; Cohort 2 = Spring 2025
- Composite Measures
  - Variables constructed using other variables (e.g., from measure items)
  - Scoring instructions, subscales, and reverse coding; syntax/programming statements
  - Scale statistics - descriptive and psychometric statistics

# 1c. Example Codebook Pages



## La Familia Table of Contents

To update: add outlining toolbar to your view click update TOC then choose Update page numbers only (only added text), or entire table (changed Heading 1 label). Please reserve Heading 1 for Scale Name use in the T of C.

Scale Name	<u>Page</u>
Instructions .....	3
Child Demographics .....	4
Mother Demographics & Migration History .....	8
Father Demographics & Migration History .....	29
Additional Variables .....	42
Community Cultural Context Rating System .....	44
Mexican American Acculturation/Enculturation Scale (MAAS).....	53
Language Use Measure.....	76
Ethnic Socialization .....	80
Family Role Models .....	88
Children's Reports of Parent Behavior Inventory .....	96
General Assessment of Relationship .....	111
Fathering Goals.....	115
Importance of Activities to Parents .....	119
Parent-Child Interactions .....	123
Parental Monitoring .....	127
Parent Adolescent Conflict Scale .....	134
Parent-Child Effective Problem Solving .....	141
Child Defiance vs. Respect for Authority.....	147
Multicultural Events Scale for Adolescents (MESA).....	153
Family Stress .....	159
Multidimensional Scale of Perceived Social Support.....	162
Acculturative Stress .....	169
Adolescent Experiences with/Perceptions of Discrimination.....	174
Parent Experiences with/Perceptions of Discrimination .....	179
Educational Values .....	185
Child Coping Strategies Checklist.....	190
Peer Delinquent Behavior.....	197
Gang Involvement .....	202
Pubertal Development Scale.....	207
Weinberger Adjustment Inventory .....	212
Economic Hardship .....	216
The Family Adaptability and Cohesion Evaluation Scales II (FACES II).....	226
Quality of Marriage Index .....	232
Resolving Conflict in Relationships (RCR).....	237

<b>Mexican American Acculturation/Enculturation Scale (MAAS) —Mother, Father, and Child Report</b>		<b>Page Last Updated: 07/28/04</b>
<b>Description</b>	<b>Associated Papers</b>	<b>Scale Directions &amp; Items</b>
<b>Item Values</b>	<b>Scoring of Scale</b>	<b>Comment</b>

### MAAS-Values - DESCRIPTION

The Mexican American Acculturation/Enculturation Scale - Values was developed by Gonzales, Knight, and Saenz. Their purpose was to develop a measure of acculturation and enculturation for Mexican Americans that is broader than and addresses the limitations of the currently available measures. The scale was developed based upon focus groups conducted with Mexican American mothers, fathers, and adolescents about the Mexican American and [Anglo American](#) cultures. Some of the items are minor re-wording of specific parent or adolescent statements made during the focus groups. Some of the items were created to reflect general topic of discussion by the focus groups. Some of the items were adapted from other measures of similar constructs.

Items 6, 16, 17, 27, 29, & 49 were taken from Sabogal et al. (1987).

The items originally read as follows (adapted for the current measure):

- 6: "Much of what a son or daughter does should be done to please the parents."
- 16: "I would help within my means if a relative told me that she/he is in financial difficulty."
- 17: "The family should consult close relatives (uncles/aunts) concerning its important decisions."
- 27: "A person should share his/her home with uncles, aunts, or first cousins if they are in need."
- 29: "One should be embarrassed about the bad things done by his/her brothers or sisters."
- 49: "One should make great sacrifices in order to guarantee a good education for his/her children."

This scale is designed to be applicable to either parent or adolescent respondents. An earlier version of the scale administered in the Bridges and Juntos projects at the PRC contained 63 items that assessed 11 subscales of values that are likely to change through the processes of acculturation and enculturation. Based on confirmatory factor analysis using Bridges data, 13 items were found not to fit into the measurement models for their respective subscales or their subscales were found to have very poor psychometric properties (i.e., egalitarian gender roles and personal freedom). These findings were replicated with Juntos data. As a result, this study dropped those 13 items and only used 50 items which assess 9 subscales of values.

# 1d. Example Codebook Pages

54

## Changes made by La Familia Group:

Item	Change	Reason
A person should be embarrassed about the bad things done by his/her relatives.	Deleted.	Problems with internal consistency based on earlier tests. Less cognitively demanding.
11,14,18,24,27,34,36,47,48	Instead of using the first person or unspecified “you” when stating a value (i.e., “You should be ready to compete with others to get ahead”), we have used the word “one”—“One should be ready to compete with others to get ahead”.	This change was made to maintain consistency in the statements and to more accurately assess general values about how people should behave/think rather than how the respondent should behave/think.
12	Changed item from “When it comes to important decisions, the family should seek advice from close relatives” to “When it comes to important decisions, the family should ask for advice from close relatives”.	Less cognitively demanding.
response format—all items	response format is now: 1) Not at all 2) A little 3) Somewhat 4) Very much 5) Completely	There were concerns with the Spanish translation of agree→disagree. The stem was changed to “how much do you believe...” in <u>order</u> to make translation easier and the cognitive task less demanding.

56

## MAAS-Values - SCALE SUBJECT INSTRUCTIONS & ITEM LIST

### Subject Instructions

*The next statements are about what people may think or believe. Remember, there are no right or wrong answers. Tell me how much you believe the following statements.*

*Las siguientes frases son acerca de lo que la gente puede pensar o creer. Recuerde, no hay respuestas correctas o incorrectas. Dígame con cuanta firmeza usted cree en las siguientes frases.*

### Father Report

Variable Name	Subscale	Item No.	Reverse Coded	Item Text
DXmas01	REL	1		Parents should teach their children to pray. Con cuánta firmeza cree que los padres deberían enseñarle a sus hijos a rezar.
DXmas02	FAMSUP	2		Parents should teach their children that the family always comes first. <u>Los</u> padres deberían enseñarle a sus hijos que la familia siempre es primero.
DXmas03	FAMOB	3		Children should be taught that it is their duty to care for their parents when their parents get old. Se les debería enseñar a los niños que es su obligación cuidar a sus padres cuando ellos envejecan.
DXmas04	FAMREF	4		Children should always do things to make their parents happy. Los niños siempre deberían hacer las cosas que hagan a sus padres felices.
				No matter what, children should always treat their

# 1e. Example Codebook Pages

73

## MAAS-Values - ITEM VALUES

NOTE: prior to recode for reverse items

Text of answer choice	Numeric value
Not at all Nada	1
A little Poquito	2
Somewhat Algo	3
Very much Bastante	4
Completely Completamente	5

## MAAS-Values - SCORING OF SCALE (provided by the Methodology Core to the team)

Mean scores for enculturation will be calculated based on the ratings of the Religion, Traditional Gender Roles, Familism, and Respect items. Mean scores for acculturation will be calculated based on the ratings of the Self-Reliance, Material Success and Competition and Personal Achievement items.

### VARIABLE LABELS FOR SCALE AND SUBSCALES

#### Enculturation Subscales:

Religion: 1, 8, 18, 27, 36, 45, 48

Traditional Gender Roles: 13, 19, 32, 42, 50

Familism: Support and Emotional Closeness: 2, 9, 20, 28, 37, 46

Familism: Obligations: 3, 11, 21, 29, 38

Familism: Family as Referent: 4, 12, 30, 39, 47

Respect: 5, 10, 15, 22, 25, 31, 40, 49

74

Religion + Traditional Gender Roles + Respect + Familism: Support & Emotional Closeness + Familism: Obligations + Familism: Family as Referent (36 items)

#### Acculturation Mean Score

Self-Reliance + Material Success + Competition & Personal Achievement (14 items)

#### Total Familism Mean Score

Familism: Support & Emotional Closeness + Familism: Obligations + Familism: Family as Referent (16 items)

#### For fathers:

dXmasrel = RELIGION (REL)

dXmasget = TRADITIONAL GENDER ROLES (GEN)

dXmasfms = FAMILISM: SUPPORT AND EMOTIONAL CLOSNESS (FAMSUP)

dXmasfmo = FAMILISM: OBLIGATIONS (FAMOB)

dXmasfmr = FAMILISM: FAMILY AS REFERENT (FAMREF)

dXmasrsp = RESPECT (RESP)

dXmaspsc = COMPETITION & PERSONAL ACHIEVEMENT (COMPPA)

dXmaspsm = MATERIAL SUCCESS (MATSUC)

dXmasisr = SELF-RELIANCE (SELFRE)

dXmasfam = FAMILISM: SUPPORT AND EMOTIONAL CLOSNESS + FAMILISM: OBLIGATIONS + FAMILISM: FAMILY AS REFERENT / 16 items

dXmasenc = RELIGION + TRADITIONAL GENDER ROLES + FAMILISM: SUPPORT AND EMOTIONAL CLOSNESS + FAMILISM: OBLIGATIONS + FAMILISM: FAMILY AS REFERENT + RESPECT / 36 items

dXmasacc = COMPETITION & PERSONAL ACHIEVEMENT + MATERIAL SUCCESS + SELF-RELIANCE / 14 items

#### For mothers:

mXmasrel = RELIGION

mXmasget = TRADITIONAL GENDER ROLES

mXmasfms = FAMILISM: SUPPORT AND EMOTIONAL CLOSNESS

mXmasfmo = FAMILISM: OBLIGATIONS

mXmasfmr = FAMILISM: FAMILY AS REFERENT

mXmasrsp = RESPECT

mXmaspsc = COMPETITION & PERSONAL ACHIEVEMENT (COMPPA)

mXmaspsm = MATERIAL SUCCESS (MATSUC)

mXmasisr = SELF-RELIANCE (SELFRE)

# Data Collection and File Creation



# Data Collection and File Creation

- How will data be input or captured?
- Data collection tools
  - Paper pencil
    - Requires data entry protocol into database
    - Best practice to perform double entry to reduce errors
  - Online tools (e.g., Qualtrics)
- Data files (spreadsheets and databases)
  - Statistical program files (e.g., R, SPSS, SAS)
    - Can embed variable meta-data
    - Can use syntax files for quality control and as documentation
  - Excel
  - Text files

# Types of Project Data and Files

- **Raw Data Files**
  - Original (item-level) data as it was collected (or entered); No changes made to data
  - Stored in data entry or collection files; may require conversion to statistical software
- **Intermediate: Merged and Processed Data Files**
  - Data that have been combined and cleaned – in this order...
    1. Within reporter
    2. Across reporters (if applicable, within cohort)
    3. Across cohorts within time
    4. Across time
  - Typically stored in statistical program file – such as SPSS –
  - Choose **one program** as the primary program for data management tasks
- **Final: Scored (Composite) Data File(s)**
  - Data that have been transformed and represent summary scores
  - Stored in file containing clean raw data and scored data (master research file)
- **Analytic File(s)**
  - Smaller files for specific analyses/manuscripts



# Key Database Elements: Quantitative Data

1. Use of ID Numbers
2. Variable Names
3. Data Formats
4. Labels
  - Variable
  - Response value codes and labels
  - Missing values

# 1a. Purpose of ID Numbers

**ID numbers**: unique identifying number

1. Method of protecting confidentiality and anonymity of participants
  - **Confidentiality**
    - Maintaining confidentiality of information collected from research participants means that only the investigator(s) or individuals of the research team can identify the responses of individual subjects; however, the researchers must make every effort to prevent anyone outside of the project from connecting individual subjects with their responses.
  - **Anonymity**
    - Providing anonymity of information collected from research participants means that either the project does not collect identifying information of individual subjects (e.g., name, address, Email address, etc.), or the project cannot link individual responses with participants' identities. A study should not collect identifying information of research participants unless it is essential to the study protocol.



# 1b. Purpose of ID Numbers

2. Uniquely identifies each record in a database
  - Each participant in a study should be assigned one reporter ID number that uniquely identifies that person that is used throughout the research project
  - Each unique type of participant / reporter / respondent
    - Students / children / adolescents
    - Parents
    - Teachers
  - Nested factors in a study will also have ID numbers in addition to reporter numbers
    - Families
    - Schools / classrooms
    - Time
3. Used to link (merge) data files

# 1c. Purpose of ID Numbers

- 4. Provides method of **linking** different types of data into a centralized project dataset in preparation for analyses
  - Provide a key unique ID number for each dataset
  - To merge, each dataset must contain key ID(s) that can be used to combine data
- To facilitate **multilevel analyses**, you must have unique IDs within each category of respondents / nesting factors based on study design
  - Teachers and students
  - Family
- For **longitudinal studies**, ID numbers are critical for linking participants' repeated measures over time
  - Cannot estimate a longitudinal model without linked data

# 1e. Example: Observational School-Based Study

- Reporters
  - Students (self-report and report on teacher)
  - Teachers (self-report and report on students)
- Time: one time point
- Nesting
  - Students nested within teachers within schools
- Three unique ID number variables

<u>SchoolID</u>	<u>TeachID</u>	<u>StudID</u>
1	101	1011
1	101	1012
1	101	1013
1	102	1021
1	102	1022
1	102	1023
1	103	1031
1	103	1032
1	103	1033
2	201	2011
2	201	2012
2	201	2013
2	202	2021
2	202	2022
2	202	2023
2	203	2031
2	203	2032
2	203	2033

## 2a. Systematic Variable Naming

- Systematic naming of variables can convey important information
  - Instrument/measure used to collect the data
  - Informant/reporter – source of data
  - Timing of data collection (e.g., wave)
  - Whether the variable was directly assessed (e.g., an item) or computed from other scores (e.g., summary, composite, or scale score)
- Variable name system should be consistent across reporters and time points
- If consistent naming conventions are used, then it is easier to ascertain the exact meaning of each variable, as well as conduct analyses.

## 2b. Variable Naming Convention Recommendations

- Develop variable naming conventions based on study design
  - Incorporate reporters / respondents
  - Incorporate time
- Have related naming conventions for different types of variables
  - Demographic variables (e.g., gender, age)
  - Item-level variables (e.g., measure / assessment items)
  - Transformed variables (e.g., reverse scored, recoded)
  - Summary-level (composite) variables (e.g., scale scores)

## 2c. Variable Naming Convention Recommendations

- Use only **8 characters** for variable names
  - Based on limitations of some statistical programs (e.g., Mplus)
  - Facilitates not having to rename variables across programs
- For longitudinal and multiple reporter data, for the same measures use the **same base variable name** with variation only by Time (Wave) or Reporter
  - E.g., Wave 1: p1ars; Wave 2: p2ars; Wave 3: p3ars
  - E.g., Parent: p1ars; Child: c1ars
  - Using the same names across reporter and times facilitates ease of syntax writing
- Examples of **BAD** variable names
  - They do not give you any information about measure/questions
  - Q1, Q2, Q245 ... etc.



# 2d. Example: Variable Naming System

Variable name character	Item-level variable names	Composite Variables (e.g., subscales, total score for a measure)	Demographic Variables
First character:	respondent type	respondent type	respondent type
Second character:	time point	time point	time point
Third character:	measure name	measure name	measure name or who is being responded about
Fourth character:	measure name	measure name	measure name
Fifth character:	measure name	measure name	measure name
Sixth character:	item number	subscale name	measure name
Seventh character:	item number	subscale name	variable type (text variable, etc.) or N/A
Eighth character:	variable type (reverse scored, recoded, text variable, etc.) or N/A	composite type (only if multiple composites are created for one measure) or N/A	N/A

## 2e. Example: Variable Names

Table 2. Example Variable Names:		
Variable Name	Variable type	Explanation
T1PTR02r	Item-level; transformed	Indicates a teacher responding at the first time point to question 2 of the PTRS, and this is the reversed scored item, not the original response.
P2BA050	Item-level; raw	Indicates a parent responding at the second time point to question 50 of the BASC, with no specific variable type.
T3PTRTO	Composite-level	Indicates a teacher's total composite score including all items on the PTRS at the third time point.
T3PTRJN	Compositve-level	Indicates a teacher's score on the Joining subscale of the PTRS at the third time point.
P1CGEN	Demographic	Indicates a parent responding at the first time point to a demographic item about their child's gender.
P1GEND	Demographic	Indicates a parent responding at the first time point to a demographic item about their own gender.
C2EV2TOr	Composite-level	Indicated a child's total raw composite score on the EVT at the second time point
C2EV2TOs	Composite-level	Indicated a child's total standardized composite score on the EVT at the second time point
C2EV2TOp	Composite-level	Indicated a child's total percentile rank composite score on the EVT at the second time point

## 3a. Data Formats: Text (String) Variables

- Open-ended, qualitative data
- Other responses: What is your ethnicity? 5 Other: Please specify\_\_\_\_\_
- If use text variables, you have to come up with coding system based on responses after data collection to code into number
- Use consistent data formats across datasets / reporters / waves

# 3b. Data Formats: Quantitative (Numeric) Data

- Before quantitative data are analyzed, the interview or survey responses must be represented by numeric codes
  - To save time and \$\$, in your data entry/collection tool, code response values as numbers and not as text to reduce errors
- Code Categories
  - Should be mutually exclusive, exhaustive, and precisely defined
  - Each response should fit into one and only one category
    - Though there are exceptions in which participants choose all that apply, such as gender, race/ethnicity
  - Ambiguity will cause coding difficulties and problems with the interpretation of the data
- Preserve Original Information
  - Code as much detail as possible
  - Recording original data, such as age and income, is more useful than collapsing or bracketing the information
    - E.g., Collect age as years: 12, 13, 14 vs. 1 = 12-15; 2 = 16-18
  - With original or detailed data, secondary analysis can determine meaningful brackets on their own rather than being restricted to those chosen by others
- Date of birth: preferable to collect as 3 numeric variables to avoid conversion problems
  - Day, Month, Year
  - Concatenate to use in calculations (during data processing)

# 4a. Variable Labels

- Variable labels are extremely important – serve as a type of documentation
  - Most statistical programs permit the user to link extended labels for each variable to the variable name
- It's helpful to provide key pieces of information:
  - (1) the item or question number in the original data collection instrument (unless the item number is part of the variable name),
  - (2) a clear indication of the variable's content,
  - (3) any research design elements that are critical (reporter, wave, etc.), and
  - (4) an indication of whether the variable is constructed from other items.
- One should develop a set of standard abbreviations in advance and present it as part of the documentation for the dataset
- Examples:
  - P2BA050 (item): “Whines”
  - C2EV2TOp (composite): “Child EVT total percentile rank – T2”
  - T3PTRTO (composite): “Teacher PTRS total score – T3”

## 4b. Response Value Labels

- Embed response value labels in your statistical dataset
  - Helps to save time and reduce errors by providing value label information
  - 0 = No; 1 = Yes
  - Likert scales: 1 = strongly disagree, 2 = disagree, 3 = agree, 4 = strongly agree

## 4b. Missing Data Codes/Labels

- Standardize missing data codes such that the same code is used for each type of missing data for all variables in and across data files
  - Be thoughtful in how you assign these codes so that they mean the same thing for all variables
- In assigning missing data codes, determine the largest variable length, then assign negative value that logically does not exist in the data
  - Just don't forget to tell your analysis software what the missing value code is
  - If you forget, by using negative values - easier to identify the problem when examining descriptive statistics (mean will be out of range)
- In general, blanks should not be used as missing data codes
  - Blanks can cause formatting problems
  - Some analytic programs require identification of missing data

# 4c. Missing Data Codes/Labels

- Reasons for missing data can be very, very important
  - Missing data because of skip patterns
  - Missing due to survey version differences
  - Missing data because of non-response (e.g., don't know; refusals)
  - Missing data because of partial survey/interview
  - Missing data due to drop-out (longitudinal studies)
- Be thoughtful in assigning what the different missing value codes in your data mean
  - Highlights importance of thorough documentation of what these codes are
  - Example: The National Longitudinal Study of Adolescent to Adult Health (Add Health)
    - Uses many different missing data codes – have to rely on documentation to know what the codes are for each variable (uses 6, 7, 8, 9, 96, 97, 98, 99, to name a few)
- Use a few codes consistently
  - E.g., -888 = don't know, -999 = skip or non-response, -777 = not applicable
- Skips could indicate the need for recoding as 0
  - Example: Do you work? No; Skips question on work hours – left blank. Might make theoretical sense to code work hours as 0



**Data  
Processing &  
Cleaning**



# Data Processing

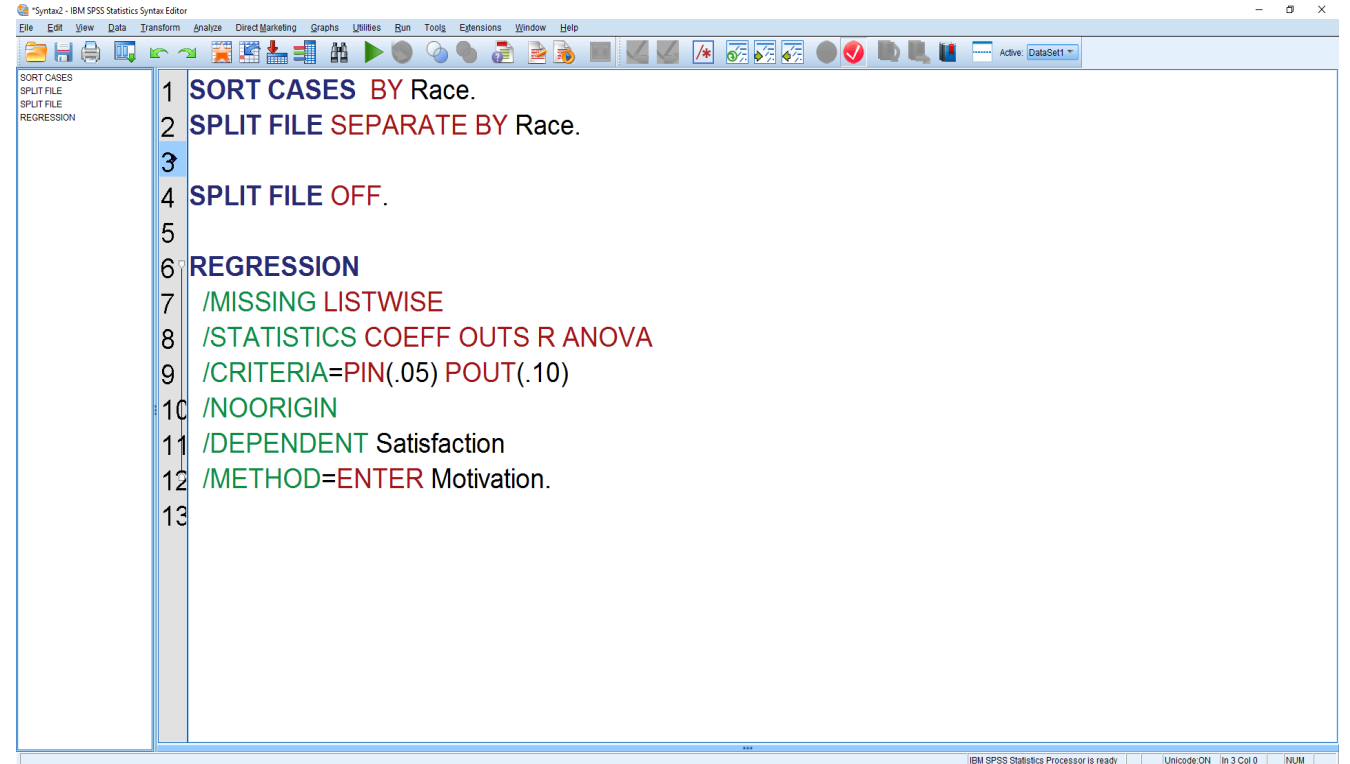
- BEFORE you begin any analyses, you will need to double-check (and ideally triple-check) that your data are accurate and formatted correctly for your analyses
- During data processing (data cleaning), careful attention to quality assurance is critical
- *“If your data are not accurate, your analyses will never be accurate.”* – Dr. Matt Fritz, every day
- To facilitate quality assurance of data processing/cleaning, syntax use is essential!



<https://monkeylearn.com/blog/data-cleaning-steps/>

# What is Syntax?

- **Syntax** (code) is a type of program for use in database and statistical programs
- Programming language of commands unique to each program (R, SPSS, SAS, Mplus, etc.)
- Type of documentation – data diary



```
1 SORT CASES BY Race.
2 SPLIT FILE SEPARATE BY Race.
3
4 SPLIT FILE OFF.
5
6 REGRESSION
7 /MISSING LISTWISE
8 /STATISTICS COEFF OUTS R ANOVA
9 /CRITERIA=PIN(.05) POUT(.10)
10 /NOORIGIN
11 /DEPENDENT Satisfaction
12 /METHOD=ENTER Motivation.
13
```

# Why use Syntax?

- Point and click (PAC) may seem easier, but using syntax has many advantages
  - Allows users to write commands that run procedures rather than using PAC
  - Allows users to perform tasks in a routine manner, reducing errors
  - Allows users to repeat tasks without recreating the wheel (reducing errors)
  - Allows users to perform tasks too tedious or difficult to do using drop-down menus
  - Allows users to document steps for future use
  - Allows users more control over analysis options
- Many programs, including SPSS, have both PAC and syntax capabilities

# Best Practices for Use of Syntax

- Document every step used in data processing including opening and saving files
  - helps you remember!
- Annotate syntax with dates and initials
  - Annotate any data management decisions that are being implemented
  - Label sections of syntax and add key words to allow for searching
- Pick a syntax convention and stick with it
  - E.g., All caps for commands
  - Indention
- In SPSS, can use drop-down menus to create syntax – don't have to memorize
  - Use **Paste** function to save commands to a syntax file

# General Steps in Data Processing/Cleaning

1. Address structural issues
2. Add variables representing design elements
3. Adherence and validity checks: Data validation
4. Creating composite scales
5. Merge data files (as needed depending on project design)
6. De-identifying data

# 1. Address structural issues

- Converting data collection / data entry file for specific statistical program(s) / analytic approach
  - Preferably export directly to statistical program file
- Confirm variables match documentation
  - Renaming variables, adding variable and value labels
  - String variables (text) have been recoded as numeric variables
  - Choose all that apply variables should be created such that there is a variable representing each response choice
    - E.g., Race (0=no; 1=yes): Var1 = White, Var2 = Black, Var3 = Asian, etc.

# 2a. Variables Representing Design Elements

- Embed variables that represent study design elements
- Why?
  - For analysis or to determine if there are different patterns of responses by design element
- Common Examples
  - Cohort
    - Group of participants on which data is collected during same historical time
    - Cohort 1: spring; Cohort 2: summer; Cohort 3: fall
  - Contact Point
    - AKA: Time, Wave, Phase
    - Timing of administration of data collection
    - Longitudinal studies; daily diary studies



## 2b. Variables Representing Design Elements

- Survey Administration Type
  - Paper pencil vs. online survey vs. interview
- Version code
  - Can represent different versions of a survey
  - E.g., different ordering
  - Different measure administration based on planned missing data design
  - Version A vs. Version B
- Completion codes (e.g., reporter, time)
  - For multiple reporter studies, completion codes for each reporter can help determine if there was partial completion within a unit (e.g., family)
  - For longitudinal studies, completion codes for each Wave can be used for attrition analyses

# 3a. Adherence and validity checks: Data validation

- Remove irrelevant data
  - Identifying bots / mischievous responders
- Record matches and counts
  - confirmation of correct number of participants and consistency across files
  - remove duplicate data
- Range and validity checking for all data
  - check for impossible and extreme scores
  - data quality through visual inspection of data
- Missing data



## 3b. Remove Irrelevant Data

- Confirm that data is from eligible, consented, enrolled participants
  - Remove ineligible or non-consented participants
  - Want the Final Research File to contain all eligible, enrolled / randomized participants
    - Do not drop/remove enrolled participants due to missing data!
    - Only remove participant data – if someone tells you they want to be removed from the study
- Online studies: Develop procedures for identifying bots/fraudulent responders
  - IP Addresses
  - Non-sensical responses to open-ended text questions
  - Include attention check questions (Choose red from the following list)
  - E.g., see Lawlor et al. (2021). *Methodological Innovations*

# 3c. Record matches and counts

1. Design a **master key file** that tracks and links study ID numbers
  - Contains all the study id numbers, identifying information, and important information that helps check id number entry errors
    - E.g., names, addresses, gender, birth date, interview dates
    - This file should be password protected for confidentiality
2. Design data collection tools to contain necessary information to confirm ID numbers
3. Confirm that all participants (ID numbers) are present and valid in the data
  - Sample size is correct based on project data (correct number of records in each data file)
  - ID numbers match across all datasets (e.g., family id number match across parent and child data)
4. Remove any duplicate cases
  - E.g., participants accidentally took survey twice – determined by exact information being entered multiple times
  - Keep the first instance of a duplicate case

# 3d. Range and validity checking for all data

- Use descriptive statistics to examine all variables
  - Do variables have the correct range (e.g., 1-5)? Cross-check with documentation.
  - Does the data make sense?
  - Does the sample generally match up with what is expected?
    - Percentage of students receiving special education supports in NE is 15%, but your sample has 55%
- If anything is 'funky' – stop and investigate
  - Figure out what is wrong and fix prior to moving on
  - Investigate extreme or impossible values
    - Example: 5-point Likert scale 1-5; but there is a score of 0
  - Using syntax, remove or fix impossible values (falls outside of the possible range for a variable)
- Reduce data entry errors by setting up data collection tools so that ranges/values are limited to acceptable values to increase data quality

## 3e. Data Quality Through Visualization

- Use data visualization (descriptive statistics) to check data quality
- Data visualization can:
  - Reveal data anomalies
  - Determine if variables have a normal distribution
  - Give the user a quick way to pinpoint areas of interest or see trends in data over time.
  - Provide insight into patterns that may arise

# 3f. Missing Data

- Develop Protocol & Procedures
  - Is partial data included in final dataset?
  - When to suspend data collection due to missing data points?
  - When to collect additional data to compensate for missing data?
  - How to code missing data?
    - How to code item-level missing data? (see prior slides on coding)
    - Missing data theory – for handling missing data in analysis (e.g., Enders, 2022)
- Validate that the amount of missing data is accurate / makes sense
  - Based on number of participants
  - Survey logic/skip patterns

# 3g. What to do when find inconsistencies or problems?

- Have a procedure for how to handle identified problems
- Examples
  - Discuss at weekly research team meetings – thoroughly document decisions in a log and in the syntax (if changing data)
  - Re-contact participant for clarification
- The raw (collected) data should **always be preserved**
- **Only** make cleaning changes on intermediary data sets using syntax that documents the reason for the change
- Document any decisions or changes that you've made on the data
  - Through syntax and other methods



# 4a. Creating Composite Variables

- **Scoring** refers to the process of creating composite scores for measures
  - **Composite (Summary) Score**: several items are combined in some way create a new score/variable
- Common types of composite scores include:
  - **Sum**: Add up the scores on the individual items (e.g., add up the number of correct items) (e.g., stressful events scale).
  - **Average**: Average the scores on the individual items (e.g., average response across 10 items each on a 5-point Likert scale).
  - **Weighted composite**: Some scales require more complex composites that weight items differently according to some formula decided on by the creator of the scale (e.g., Minnesota Multiphasic Personality Inventory).
- **Scoring instructions** should be based on the original source of the measure
  - E.g., mean of items 1-5
  - E.g., reverse score item 1, then create a mean of items 1-6
  - E.g., sum all items
  - E.g., count number of yes responses to create an index score

# 4b. Scoring Recommendations

- Never recode into the same variable
  - When a variable means one thing at a stage of a project, and another at another stage, it can really mess things up
  - **Always preserve original data**
- With dummy variables, consider putting what a value of 1 means in the variable name or label so you know at a glance what direction it is coded
- Reverse score (as necessary) prior to creating composite
- When creating new variables, attach variable labels
- Always run a sanity check on a constructed variable
  - Always test untried commands
  - Build in checks on mathematically impossible transformations
  - Does the new score make sense?



# 5. Merging Data

- This is only applicable if you have multiple datasets for one study
  - Examples: Multiple reporters, waves, cohorts, types of data, etc.
- Multiple time points or reporters:
  - Requires the use of ID numbers that can be used to combine data files
- Multiple cohorts:
  - Requires the same variables that are combined have to have the exact same properties across files
- Example
  - Step 1: Merge mother T1 & child T1
  - Step 2: Merge mother T2 & child T2
  - Step 3: Merge T1 (all reporters) & T2 (all reporters) to get final dataset
- After merging files – confirm that the number of variables and participants is correct

# 6. De-identify Data

- Identifying information: In quantitative research, the identity of the individual participants is rarely relevant to the research purpose.
- Create data de-identification rules and procedures (how and when to de-identify for each data type)
  - Ideally, identifying information should be stored in a separate/file location and an anonymous case identifier should be used to link the identifying information with the participant's other data
  - May collect some identifying information as part of a survey. This information needs to be removed prior to sharing data or closing out a study
- Identifiers to consider for removal (if someone can be identified from a combination of information)
  - Names
  - Geographic subdivisions smaller than state (addresses)
  - Birth date
  - Phone numbers
  - Email
  - Social security numbers
  - Medical record numbers
  - IP address numbers
  - Biometric identifies
  - Images

A photograph of a server room. The foreground shows several server racks with blue indicator lights. The background is filled with more server racks and a bokeh effect of yellow and white lights, suggesting a large, active data center.

# Data Storage and Access

# Data Storage and Access Protocols

- In creating rules, consider relevant data uses and security needs
  - Confidentiality, frequency of collection, ease of access to the data, staff needed to make data accessible, type of analysis
  - Secure parameters for data transfer and storage (e.g., video upload procedures)
- Rules for each collection method
  - Qualtrics (e.g., survey ownership)
  - Website (e.g., how are data accessed; log-in information)

# 1. Data Storage

- Consider each storage type (physical filing cabinets, websites, OneDrive, internal server space, etc.)
- Consider needed security (e.g., identifiable data vs. de-identifiable data)
- Document describing where data are stored and the rationale for location
- Consider miscellaneous data in addition to survey and assessment data
  - Fidelity
  - Record reviews
  - Publicly-available data
  - Data requests (e.g., NDE, school district, attendance)

# 1a. Electronic Data Storage

- Develop organizational structure for storing all project files
  - Centralized location for project team
  - Acts as a type of documentation
- Electronic data folder system to store files
  - Subfolders for programs, data sets, analyses, documentation
  - Reflects various types of data and the timing or waves of data collection
  - Have naming convention for folders, data files, syntax files
    - How will files be named? What naming conventions will be used to achieve consistency?



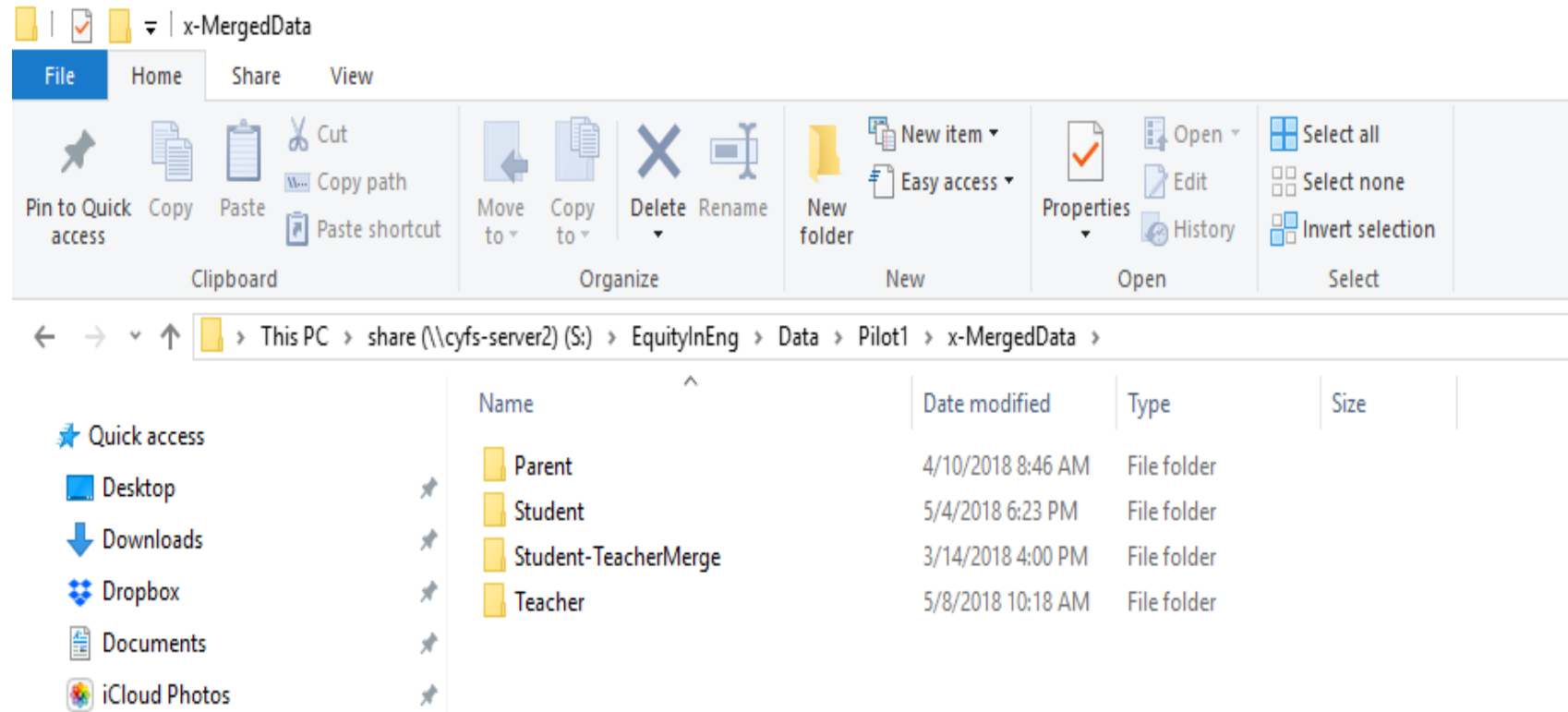
# 1b. Example: File Structure

The screenshot shows a Windows File Explorer window with the following details:

- Address Bar:** This PC > share (\\cyfs-server2) (S:) > EquityInEng
- Command Bar:** File, Home, Share, View. Ribbon tabs include Clipboard (Pin to Quick access, Copy, Paste, Copy path, Paste shortcut), Organize (Move to, Copy to, Delete, Rename), New (New folder, New item, Easy access), Open (Properties, Open, Edit, History), and Select (Select all, Select none, Invert selection).
- Left Pane (Navigation):** Quick access (Desktop, Downloads, Dropbox, Documents, iCloud Photos, Pictures, iCloud Drive, Codebooks, Lectures).
- Main Pane (File List):**

Name	Date modified	Type	Size
Conferences_Presentations	3/30/2018 10:31 AM	File folder	
Data	5/10/2018 2:21 PM	File folder	
Documentation	5/7/2018 3:31 PM	File folder	
Literature	3/19/2018 1:35 PM	File folder	
Manuscripts	5/11/2018 12:49 PM	File folder	
Staff	4/11/2018 11:32 AM	File folder	

# 1c. Example: File Structure



## 2. Data Access Rules

- Determine data access rules for data depending on level of needed security (e.g., PII data including videos, recruitment information vs de-identified survey data files)
  - Scope of data to be shared
  - How to request access?
  - Who approves access?
  - How will data be shared?
  - Specify plans for those outside of UNL.
    - Consider need for data use transfer agreement/data sharing agreement



# Resources

# Citations

- 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>
- Enders, C. K. (2022). *Applied missing data analysis*. Guilford Publications.
- Lawlor, J., Thomas, C., Guhin, A. T., Kenyon, K., Lerner, M. D., Ucas Consortium, & Drahota, A. (2021). Suspicious and fraudulent online survey participation: Introducing the REAL framework. *Methodological Innovations*, 14(3), 20597991211050467.
- Levesque, R. J. R. (2017). Data Sharing Mandates, Developmental Science, and Responsibly Supporting Authors. *Journal of Youth and Adolescence*, 46, 2401-2406.
- Roosa, M. W., Liu, F. F., Torres, M., Gonzales, N. A., Knight, G. P., & Saenz, D. (2008). Sampling and recruitment in studies of cultural influences on adjustment: a case study with Mexican Americans. *Journal of family psychology*, 22(2), 293-202.
- Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, 61, 726-728.



# Local Resources

- UNL Libraries
  - Research Partnerships
  - <https://unl.libguides.com/RP>
  - <https://libraries.unl.edu/research-data-management/>
  - [https://libraries.unl.edu/images/Services/Data\\_management\\_plan\\_template.pdf](https://libraries.unl.edu/images/Services/Data_management_plan_template.pdf)
  - <https://unl.libguides.com/datamanagement/data-management-plans>
- UNL Research Data Strategy Task Force
  - <https://researchdata.unl.edu/>
- UNL Research Core Facilities
  - <https://research.unl.edu/proposaldevelopment/core-facilities/>
  - MAP Academy Advanced Analytics & Data Infrastructure Core
  - [Mapacademy.unl.edu](http://Mapacademy.unl.edu)



# Resources

- Data Management and Sharing Plans Tool: <https://dmptool.org/>
- Inter-university Consortium for Political and Social Research's (ICPSR) Guide to Social Science Data Preparation and Archiving: <https://www.icpsr.umich.edu/files/deposit/dataprep.pdf>
- NIH Data Management and Sharing Policy: <https://osp.od.nih.gov/scientific-sharing/nih-data-management-and-sharing-activities-related-to-public-access-and-open-science/>
- NIA Data management and sharing requirements: Tips and tricks to plan ahead: <https://www.nia.nih.gov/research/blog/2021/03/data-management-and-sharing-requirements-tips-and-tricks-plan-ahead>
- Registry of Research Data Repositories: <https://www.re3data.org/>
- U.S. Department of Health and Human Services Methods for De-identification: <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html>



# Fair Data Principles

- <https://www.go-fair.org/fair-principles/>
- Wilkinson et al (2016). *Scientific Data*.

## FAIR Principles

## Compliance



### Findability

Resource and its metadata are easy to find by both, humans and computer systems. Basic machine readable descriptive metadata allows the discovery of interesting data sets and services.

- ✓ F1. Resource is uploaded to a public repository.
- ✓ F2. Metadata are assigned a globally unique and persistent identifier.



### Accessibility

Resource and metadata are stored for the long term such that they can be easily accessed and downloaded or locally used by humans and ideally also machines using standard communication protocols.

- ✓ A1. Resource is accessible for download or manipulation by humans and is ideally also machine readable.
- ✓ A2. Publications and data repositories have contingency plans to assure that metadata remain accessible, even when the resource or the repository are no longer available.



### Interoperability

Metadata should be ready to be exchanged, interpreted and combined in a (semi)automated way with other data sets by humans as well as computer systems.

- ✓ I1. Resource is uploaded to a repository that is interoperable with other platforms.
- ✓ I2. Repository meta- data schema maps to or implements the CG Core metadata schema.
- ✓ I3. Metadata use standard vocabularies and/or ontologies.



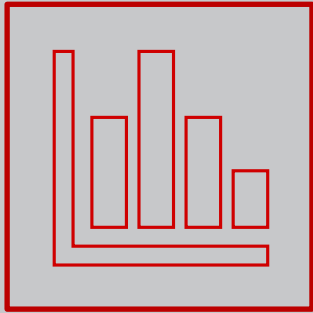
### Reusability

Data and metadata are sufficiently well-described to allow data to be reused in future research, allowing for integration with other compatible data sources. Proper citation must be facilitated, and the conditions under which the data can be used should be clear to machines and humans.

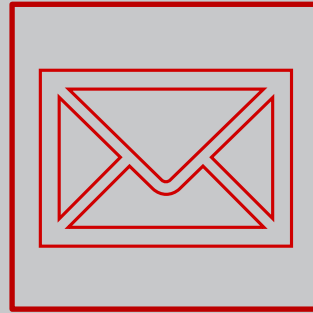
- ✓ R1. Metadata are released with a clear and accessible usage license.
- ✓ R2. Metadata about data and datasets are richly described with a plurality of accurate and relevant attributes.



# Thank you! Questions and Discussion



Supported by Nebraska Research Initiative Funding to the Advanced Analytics and Data Infrastructure Core in the MAP Academy



Lorey A. Wheeler  
Director, MAP Academy  
Advanced Analytics & Data Infrastructure Core  
[lorey@unl.edu](mailto:lorey@unl.edu)



[mapacademy.unl.edu](http://mapacademy.unl.edu)