Pursue what's possible.

Data Literacy & Management

for EVERYONE

Data Services 2022

Pursue what's possible.

Meet the Data Services Team:

Jennifer Moore
Head of Data Service

Sarah Swanz
Humanities Data Curator & Data Services Librarian

Dorris Scott
Social Science Data Curator & GIS Librarian

Mollie Webb
Data & GIS Developer

Bill Winston
GIS & Data Visualization Analyst

Core Services:

- Data Management
- Data Curation & Sharing
- Data Literacy
- Data Analysis
- Data Visualization
- Geographic Information Systems (GIS)

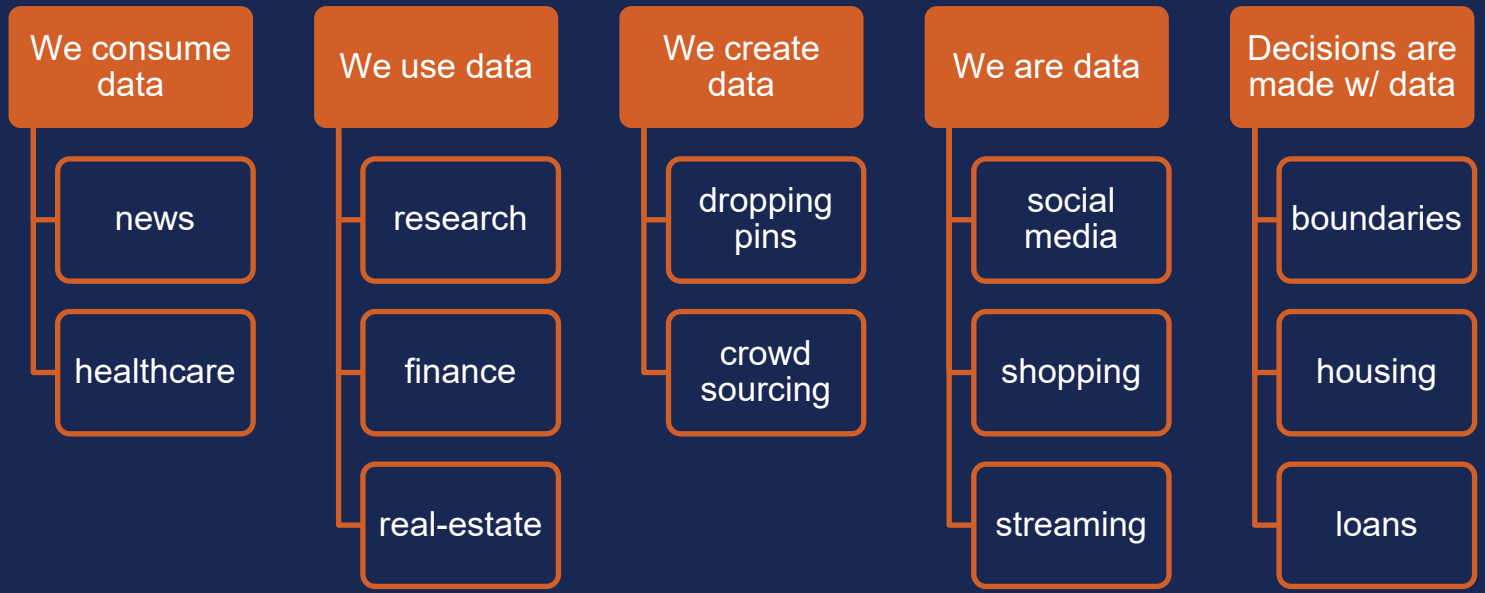
Outline

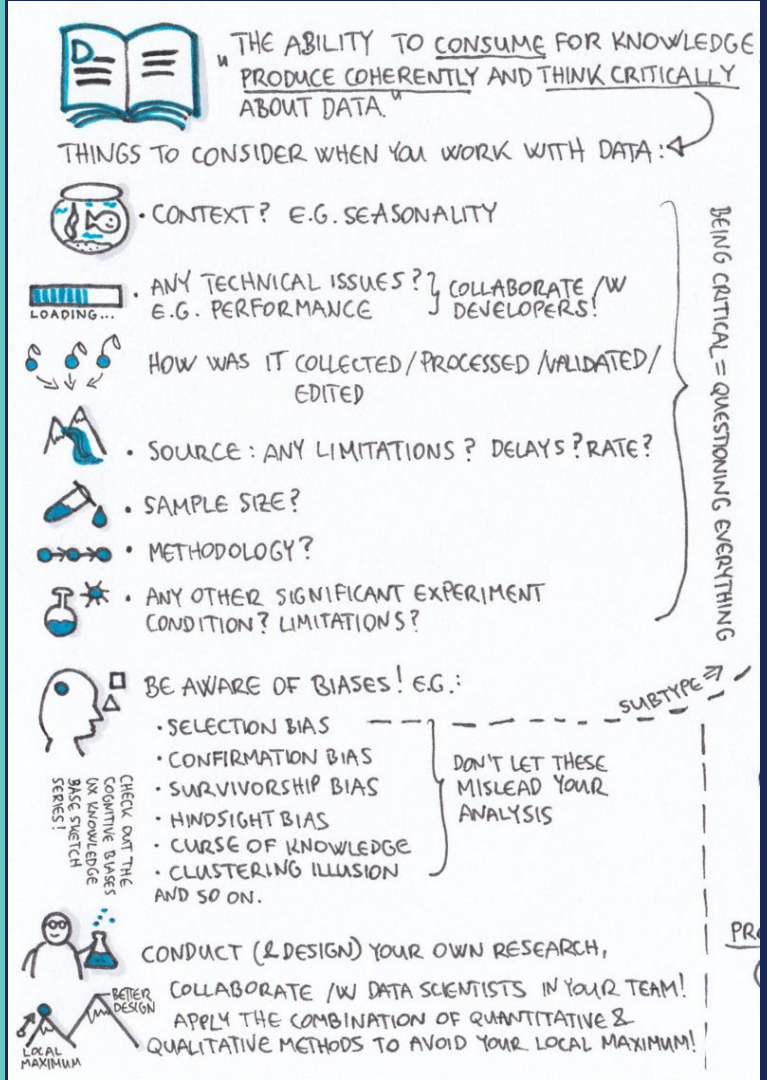
- What we mean by data
- Why you should care
  - Trusting Data (data lit)
  - Understanding Data (data man)
- What EVERYONE needs to know (data lit)
- Practices EVERYONE can adopt (data mang)

What we mean by…..

Term	Definition
Data	Observable facts, measurements, statistics, behaviors, other phenomena, record in a variant of formats
Data literacy	Ability to find, evaluate, understand, and analyze, use data to make convincing arguments, assess the arguments of others, and understand the data lifecycle
Data management	Organizing data in a consistent and logical way so that is findable, accessible, and reusable.

Why should we care



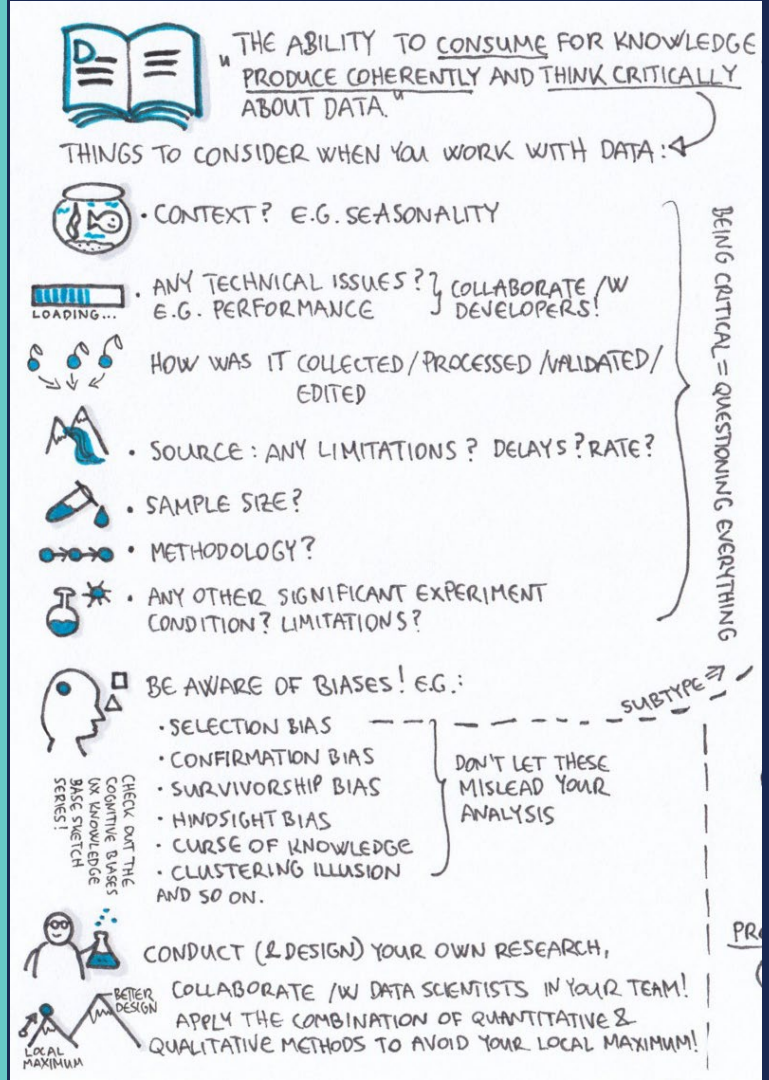


Pursue what's possible.

What EVERYONE needs to know about data

On the surface considerations...

- What is the data source?
- Absolute or proportional values?
- What is the margin of error?



Pursue what's possible.

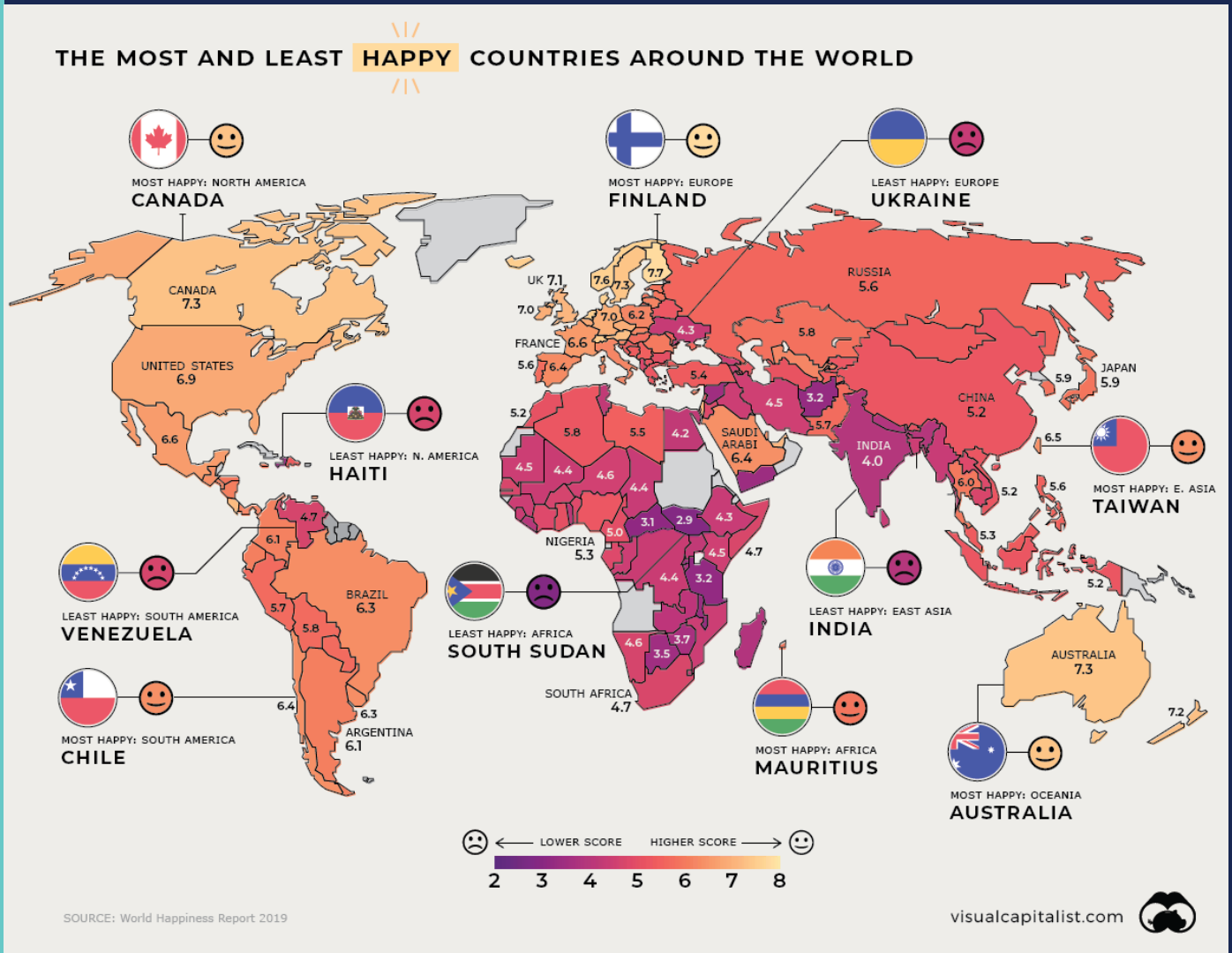
What EVERYONE needs to know about data

Under the surface considerations...

- How was the data collected?
- What is the sample size?
- What biases are present?

Pursue what's possible.

Test your skills!



World's Most & Least Happiest Countries

Pursue what's possible.

2019 Apps by Worldwide Downloads

Pursue what's possible.

Data Management

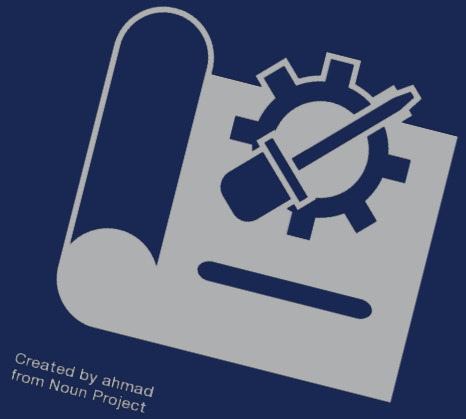
It's all about a plan

Planning

- **For funding:** The DMPTool – create a plan for specific funders using boilerplate language from WashU.
- **For everyday:** Keep in simple

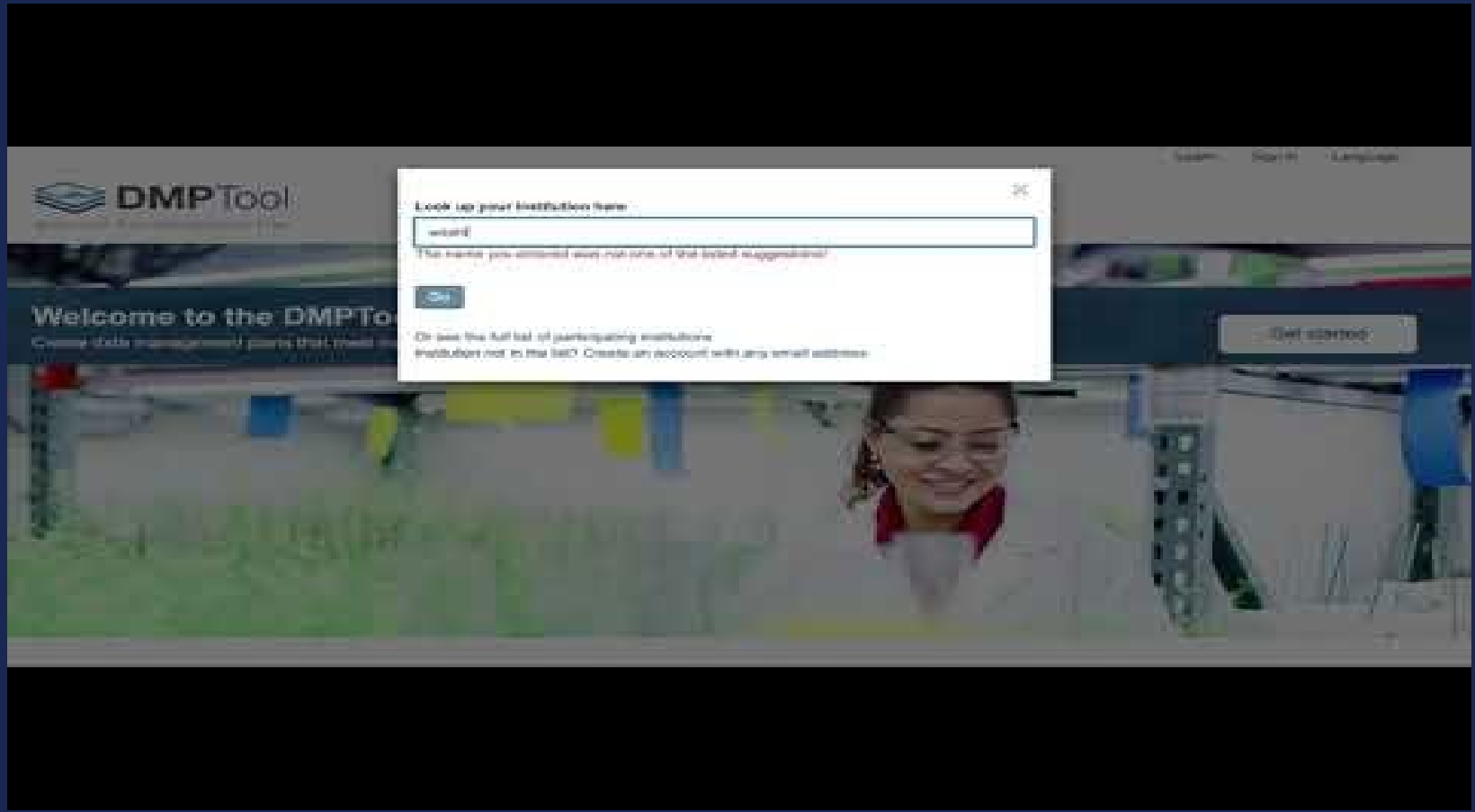
Funders: data management plan

- What will you collect?
- How will you store it?
- How will organize it?
- How will you collaborate?
- How will you document it?
- How will you protect it?
- How will you share it?



Data Literacy & Management for EVERYONE

Pursue what's possible.



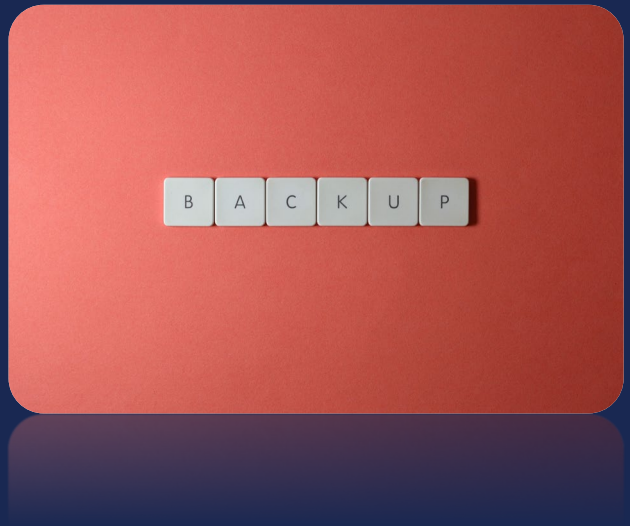
Practices EVERYONE can adopt

1. Storage
2. Organization
3. Consistency
4. Documentation

Pursue what's possible.

1. Storage

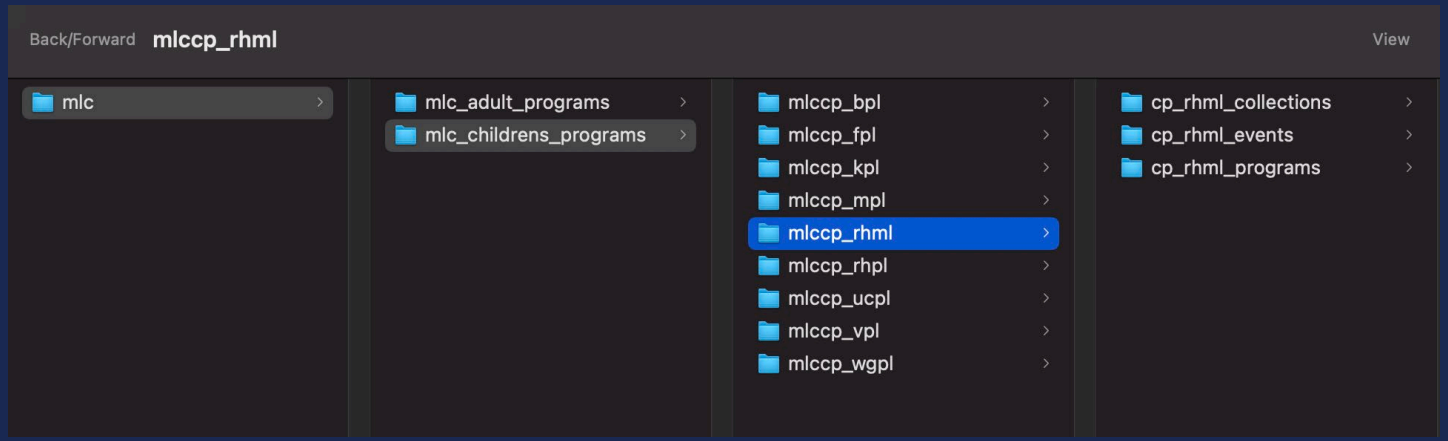
- YOUR COMPUTER IS NOT SAFE ENOUGH
- Always have at least two locations
- Cloud storage is a good option



2. Organization

1. Hierarchy is helpful, but it's important not to get too deep (3-5 levels is ideal)
2. Avoid overlapping categories
3. Folder names should be short and meaningful
4. Do not rely on nested folder structures
5. In a non-hierarchical structure, you can use tags, but these should be thoughtful and consistent.

File hierarchy's aid collaboration

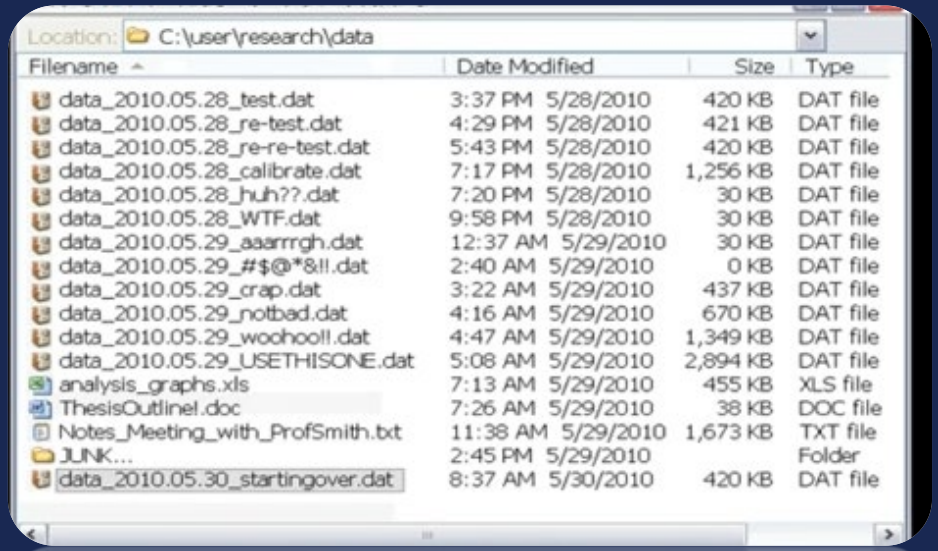


3. Consistency: File Naming Best Practices

BRIEF (32c max) but MEANINGFUL
Don't rely on nested folders
Use consistent structure
Use dates in YYYYMMDD format
List versions alpha-numerically
NO SPACES!
NO SPECIAL CHARACTERS! (#$%@.*^....)

File naming convention keeps things consistent and findable

What's Wrong Here?

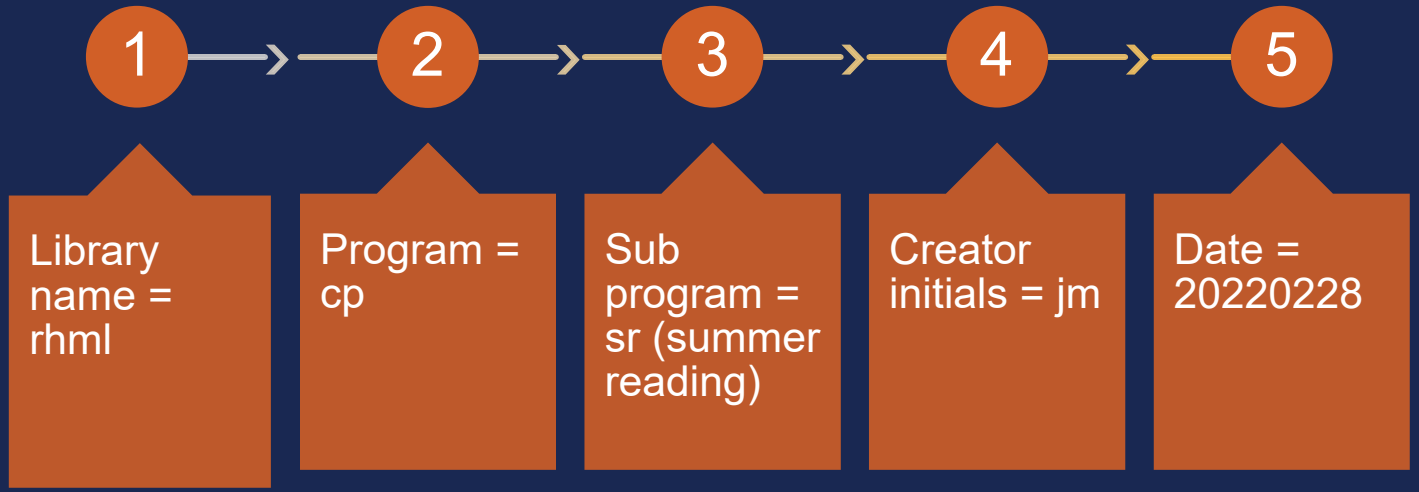


Example: US Census

ACS_16_5YR_B1006

US Census Convention

From the MLC example



rhml_cp_sr_jm_20220228

Use Formatted Tables

Keep	Parse	Form	UID	Standardize
•Don't edit the originals •Duplicate and work from the duplicate	•1 cell = 1 data type/level	•Headers: •No spaces •No special characters •Short, but meaningful •Start with letter	•Auto generated •001, 002, 003 •Combined •US Census Tracts •MO = 29 •STL City = 510 •Tract = 2104 **•ID = 295102104**	•use data validation •avoid variants, spelling mistakes •identify expected values •use rules

Table Example

program_id	year	theme	particiapants	lead_by
xxx-xxxx-xx	2001	name1	5	librarian name
xxx-xxxx-xx	2002	name2	10	librarian name
xxx-xxxx-xx	2003	name3	15	librarian name
xxx-xxxx-xx	2004	name4	20	librarian name
xxx-xxxx-xx	2005	name5	25	librarian name
xxx-xxxx-xx	2006	name6	30	librarian name
xxx-xxxx-xx	2007	name7	35	librarian name
xxx-xxxx-xx	2008	name8	40	librarian name
xxx-xxxx-xx	2009	name9	45	librarian name
xxx-xxxx-xx	2010	name10	50	librarian name
xxx-xxxx-xx	2011	name11	55	librarian name
xxx-xxxx-xx	2012	name12	60	librarian name
xxx-xxxx-xx	2013	name13	65	librarian name
xxx-xxxx-xx	2014	name14	70	librarian name
xxx-xxxx-xx	2015	name15	75	librarian name
xxx-xxxx-xx	2016	name16	80	librarian name
xxx-xxxx-xx	2017	name17	85	librarian name
xxx-xxxx-xx	2018	name18	90	librarian name
xxx-xxxx-xx	2019	name19	95	librarian name
xxx-xxxx-xx	2020	name20	100	librarian name
xxx-xxxx-xx	2021	name21	105	librarian name
xxx-xxxx-xx	2022	name22	110	librarian name
xxx-xxxx-xx	2023	name23	115	librarian name
xxx-xxxx-xx	2024	name24	120	librarian name
xxx-xxxx-xx	2025	name25	125	librarian name

Reusing Tables

- ✓ Make sure the source data is well-documented and has licensing information
- ✓ Interrogate the dataset for issues and limitations
- ✓ Keep a copy of the data, untouched
- ✓ Clean your copy so it conforms to best practices
- ✓ Document changes
- ✓ Give attribution to source

Pursue what's possible.

Basic documentation

This codebook.txt file was generated on <YYYYMMDD> by <Name>

GENERAL INFORMATION
---

1. Title
2. Author Information
3. Date
4. Contextual description of the data

FILE OVERVIEW
---

1. File List
2. Relationship between files:
3. Additional related documents
4. Are there multiple versions of the dataset? yes/no

METHODOLOGICAL INFORMATION

Description of methods used for collection/generation of data:

Pursue what's possible.

Questions