Data Literacy & Management

for EVERYONE
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…..

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>Observable facts, measurements, statistics, behaviors, other phenomena, record in a variant of formats</td>
</tr>
<tr>
<td>Data literacy</td>
<td>Ability to find, evaluate, understand, and analyze, use data to make convincing arguments, assess the arguments of others, and understand the data lifecycle</td>
</tr>
<tr>
<td>Data management</td>
<td>Organizing data in a consistent and logical way so that is findable, accessible, and reusable.</td>
</tr>
</tbody>
</table>
Why should we care

We consume data
- news
- healthcare

We use data
- research
- finance
- real-estate

We create data
- dropping pins
- crowd sourcing

We are data
- social media
- shopping
- streaming

Decisions are made w/ data
- boundaries
- housing
- loans
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?
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?
Test your skills!
2019 Apps by Worldwide Downloads

Top 5 New Apps:
- MV Master
- Disney+
- Jianying Vlog
- XRecorder
- Stickerly

Top 5 Publishers:
- Facebook
- Google
- ByteDance
- Voodoo
- Tencent

Source:Sensor Tower
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 it 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
Practices EVERYONE can adopt

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

1. Type of survey = American Community Survey
2. Date = year covered
3. Range = estimate
4. Specific table = table id

ACS_16_5YR_B1006
From the MLC example

- Library name = rhml
- Program = cp
- Sub program = sr (summer reading)
- Creator initials = jm
- Date = 20220228

rhml_cp_sr_jm_20220228
Use Formatted Tables

**Keep**
- Don’t edit the originals
- Duplicate and work from the duplicate

**Parse**
- 1 cell = 1 data type/level

**Form**
- Headers:
  - No spaces
  - No special characters
  - Short, but meaningful
  - Start with letter

**UID**
- Auto generated
- 001, 002, 003
- Combined
- US Census Tracts
  - MO = 29
  - STL City = 510
  - Tract = 2104
  - ID = 295102104

**Standardize**
- use data validation
- avoid variants, spelling mistakes
- identify expected values
- use rules
### Table Example

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