A very “Meta-” Data Management Training Workshop

Lisa Johnston (ljohnsto@umn.edu)
University of Minnesota - Twin Cities
Midwest Data Librarian Symposium 2015
Agenda

9:00 - 9:20   Icebreaker
9:20 - 9:30   Lecture - Backwards Design (Lisa)
9:30 – 11:50  Show and Tell
    Round 1 (#1-9) Starts at - 9:30
    Q&A + Discussion
    Round 2 (#10-18) Starts at – 10:05
    Q&A + Discussion

Break 10:30

    Round 3 (#19-27) Starts at - 10:40
    Q&A + Discussion
    Round 4 (#28-36) Starts at 11:15
    Q&A + Discussion

11:50 - Wrap up discussion on DIL competencies
12:00 End of Workshop
Icebreaker
Also some great tools to use in your teaching!
Backwards Design -
Rules to live by when creating educational opportunities
Backwards Design

Define Goals
What do you want your students to know or be able to do as a result of this workshop/learning moment?

Define Acceptable Evidence
How will you know students have achieved the goals?

Create Instruction & Learning Activities
How will you provide instruction and practice so students can do well on that assessment?
Example backwards design with this workshop

Librarian Competency: Teach/educate on Data Information Literacy Topics

- **Goal/Learning Outcome:** All participants will leave this session with exciting new ideas and examples to use in future teaching opportunities.

- **Acceptable Evidence:** Students are actively discussing the techniques presented by their peers and ask questions/offer feedback.

- **Instruction/learning activity:** Each participant will share one "highlight" from their own data management teaching toolkit. There will be interval breaks for feedback and group discussion.

- **Implement and Reflect:** In progress!
Side Bar -
Data Information Literacy Competencies
12 DIL Competencies

<table>
<thead>
<tr>
<th>Data Processing and Analysis</th>
<th>Data Curation and Re-Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Management and Organization</td>
<td>Data Conversion and Interoperability</td>
</tr>
<tr>
<td>Data Preservation</td>
<td>Data Visualization and Representation</td>
</tr>
<tr>
<td>Databases and Data Formats</td>
<td>Discovery and Acquisition</td>
</tr>
<tr>
<td>Ethics and Attribution</td>
<td>Metadata and Data Description</td>
</tr>
<tr>
<td>Data Quality and Documentation</td>
<td>Cultures of Practice</td>
</tr>
</tbody>
</table>

DIL Project: Competencies Ranked by Fac/Students (2012-2013)
Show-and-Tell
Data management training highlights from your peers!
Data Sharing Pro/Con Exercise
Lisa Johnston
University of Minnesota Libraries
# Data Sharing Techniques

<table>
<thead>
<tr>
<th>Ways to Share your Data</th>
<th>Pros?</th>
<th>Cons?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post online to a personal or project website</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publish data in a journal as a “supplement” to your main research article.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Make your data “Available on request” via email or dropbox to those who ask.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deposit in a disciplinary repository (e.g. Dryad, FlyBase, etc.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deposit in an institutional repository, such as the Data Repository for the University of Minnesota (DRUM)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lisa Johnston, University of Minnesota Libraries (ljohnsto@umn.edu)
Flippin’ DMPTool Instruction
Amanda K Rinehart
The Ohio State University Libraries
Flippin’ DMPTool Instruction: type first, instruct later

1) Provide participants with:
   Case studies (for those without their own research details)
   Handout of resources
   Expectation of leaving with a DMP outline

2) Log them into the DMPTool, get them to the template page

3) Answer/collect DIL questions as the class works

4) Review questions and discuss challenges at the end

Results: Active and self-reflective participants, product-oriented session
#3

Bingo. Candy. DMPs.

Megan N. O'Donnell
Iowa State University Library
mno@iastate.edu / @Mega_NO

Teaching Materials: DMP Bingo - the good, the bad, the ugly. Figshare. http://dx.doi.org/10.6084/m9.figshare.1564825
Outcomes from 1st Deployment

Candy breaks the ice.

Playing the game leads to questions and discussions – among groups and as a class.

Working with “real” DMPs presents challenges that “canned” DMPs could avoid.

In exchange, they have gravitas...

Vocabulary is more difficult than concepts.
#4
Data Repositories for all!
Sharing and finding data sets

Michelle Bass
University of Chicago, The John Crerar Library
Starting Points

REGISTRIES

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

biosharing.org
Information Resources

COMPiled LISTS

Special Library Association,
Physics-Astronomy-Mathematics
Division

Scientific Data
Who Owns Your Data: Identifying Policies

Abigail Goben
University of Illinois at Chicago
## Data Policies

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>My Lab or Research Team</td>
<td></td>
<td></td>
</tr>
<tr>
<td>My College: What does the Dean of Research Say?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>My Institution: Check with Office of Research, Sponsored Programs, or Board of Trustees. (Hint, it may be in the IP Policy)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>My Funder: Do they retain ownership? Private or Government?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>My Journal: Do they require or encourage deposit? If so, under what license?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abigail Goben, University of Illinois at Chicago (agoben@uic.edu)
#6
Researcher Generated Metadata in a Self-Submission IR model

Amy Koshoffer
University of Cincinnati Libraries
View a “good” record In Scholar@uc

Discuss metadata

Give the participants a description of the experiment + a blank form

Go through the process of pulling out metadata from description

Compare to what is actually in Scholar

Amy Koshoffer, University of Cincinnati Libraries (koshofae@uc.edu)
Data Horror Story & Data Best Practice

Kristin Briney
University of Wisconsin-Milwaukee
Heroic Computer Dies To Save World From Master's Thesis

WALTHAM, MA—A courageous young notebook computer committed a fatal, self-inflicted execution error late Sunday night, selflessly giving its own life so that professors, academic advisors, classmates, and even future generations of college students would never have to read Jill Samoskevich’s 227-page master’s thesis, sources close to the Brandeis University English graduate student reported Monday.

"This fearless little machine saved me from unspoken hours of exasperated head-scratching and eyestrain, as well as years of agonizing self-doubt over my decision to devote my life to teaching," said professor John Rebson, who had already read through three drafts of Samoskevich’s sprawling, 38,000-word dissertation, titled A Hermeneutical Exploration Of Onomatopoeia In The Works Of William Carlos Williams As It May Or May Not
Follow the 3-2-1 Rule

3 copies of your data

In 2 different locations

On more than 1 type of storage hardware
#8
Identifying data need in context of behavior and workflow

Tina Griffin
University of Illinois at Chicago
Initial Consultation: Individ. or Group worksheet

Data management group worksheet basic science:
Describe your research question in two or so sentences. Note if you are working on this with others:

How do you organize your data?
Examples: chronologically, assay type, project type

What type of project is this: Main, Pilot, Collaboration, Publication follow-up, Grant prep, Other

What documentation do you struggle with?

What stage is this project in: Beginning, Middle, End

Why do you struggle with it?

List the type of assays associate with this project and the output:
Example: Immunohistochemistry – image file, [or] protein quantification – Excel

What lab-wide changes in data management would you like to see?

For each assay type, where is the data stored (note paper [P] vs digital [D] formats)?

How long does it take you to document your research work?

For each assay type, where is the supporting information stored (reagents, protocols, animal info, etc.)?

How often do you get behind?

For each assay type, what information is essential to record?

How long does it take you to catch up?

Do you have any naming conventions Y/N? If yes, give a few examples of what they look like:

How long do you think documentation should take?

What do you want your PI to help you with regarding data management?

Are you willing to spend more time on data management in order to keep track of your work? Y/N

How much longer are you willing to spend?

What do you want your librarian to help you with regarding data management?
#9

Summary of New Federal Guidelines

Cunera Buys
Northwestern University
Agency Responses Summary - Articles

AGENCIES USING PUBMEDCENTRAL
Agency for Healthcare Research and Quality (AHRQ)
HHS Office of the Assistant Secretary for Preparedness and Response (ASPR)
Centers for Disease Control and Prevention (CDC)
Food and Drug Administration (FDA)
National Aeronautics and Space Administration (NASA)
National Institutes of Health (NIH)
National Institute of Standards and Technology (NIST)
United States Department of Veterans Affairs (VA)

AGENCIES USING DOE’S PAGES (Public Access Gateway for Energy & Science)
Department of Energy (DOE)
National Science Foundation (NSF)

AGENCIES WITH OWN REPOSITORIES
Department of Defense (DOD)--Defense Technical Info Center
National Oceanic and Atmospheric Administration (NOAA)
United States Department of Agriculture (USDA)-USDA public access archive system

OTHER (TBD)
Department of Transportation (DOT)
United States Agency for International Development (USAID)
United States Geological Survey (USGS)

Cunera Buys, Northwestern University
Agency Responses Summary
Time Frame for Depositing Data in a Publically Accessible Repository

At time of article publication
Agency for Healthcare Research and Quality (AHRQ)
Department of Energy (DOE)
Food and Drug Administration (FDA)
National Institutes of Health (NIH)
National Institute of Standards and Technology (NIST)
National Science Foundation (NSF) (exploring this option)
United States Agency for International Development (USAID)

With article publication or within 30 months of collection
HHS Office of the Assistant Secretary for Preparedness and Response (ASPR)
Centers for Disease Control and Prevention (CDC)

With article publication or within 1 year of collection
National Oceanic and Atmospheric Administration (NOAA)

At time of publication or within a reasonable time period after publication
National Aeronautics and Space Administration (NASA)

Within a reasonable time
Department of Defense (DOD)-- Defense Technical Info Center

Doesn’t specify
United States Department of Veterans Affairs (VA)
United States Department of Agriculture (USDA)
Department of Transportation (DOT)
United States Geological Survey (USGS)
Thinking critically about data visualizations

Cynthia Hudson-Vitale
Washington University in St. Louis
Graphs/visualization considerations

The following things are important to consider when looking at a graph:

1. **Title**
2. Labels on both axes of a line or bar chart and on all sections of a pie chart
3. Source of the data
4. Key to a pictograph
5. Uniform size of a symbol in a pictograph
6. Scale: Does it start with zero? If not, is there a break shown
7. Scale: Are the numbers equally spaced?

Cynthia Hudson-Vitale, Washington University in St. Louis (chudson@wustl.edu)
#10
Thinking critically about data visualizations

Graphs/Infographics in action --

1. choose an image at http://badgraphs.tumblr.com or visual.ly

2. apply considerations and determine if it is misleading – in what manner? – how might it be improved?

3. report back

Cynthia Hudson-Vitale, Washington University in St. Louis
(chudson@wustl.edu)
# 11
Data Memes to Maximize Student Amusement!
Brianna Marshall
Research Data Services, University of Wisconsin-Madison
Data Memes to Maximize Student Amusement!

See it in action: https://speakerdeck.com/bmarshall/research-data-management-in-five-easy-ish-steps

Brianna Marshall, University of Wisconsin-Madison (brianna.marshall@wisc.edu)
Mapping data outcomes for better planning

Heather Coates
Indiana University-Purdue University Indianapolis
<table>
<thead>
<tr>
<th>Research Questions</th>
<th>Analyses</th>
<th>Variables, Data Helds</th>
<th>Source/Collection Method</th>
<th>Data Requirements &amp; Assumptions</th>
<th>Analytical Notes</th>
<th>Potential Collection Issues</th>
<th>Potential Processing Issues</th>
<th>Quality Standards</th>
<th>QA/QC Procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does body image differ between racial/ethnic groups?</td>
<td>chi square</td>
<td>Q8, COMPARE COMBINED-COMPUTED, Q8 YOURTHOUGHTSONYOURBODY</td>
<td>World Health Organization, Health Behaviors of School-aged Children</td>
<td>1. Each person, item, entity is included in only one cell of the contingency table (i.e., no repeated measures). 2. Expected frequencies should be greater than 5</td>
<td>Look at row and column percentages to interpret any effects</td>
<td>Identify response rate required for sufficient power; Does sampling method address this? Time required to complete, motivation, validity of self-reported data. Prevent missing responses</td>
<td>How to handle missing and inconsistent data? Need to identify multiple types of missing data? How to handle logical errors?</td>
<td>Valid range, Expected distribution, Expected missing cases</td>
<td>Screen with frequency distributions and check standard deviation for sample and by racial/ethnic groups. Check missing cases for variables against overall missing data rate. Run normality tests (K-S or S-W). Screen with frequency distributions and check standard deviation for sample and by racial/ethnic groups. Check missing cases for variables against overall missing data rate. Run normality tests (K-S or S-W).</td>
</tr>
<tr>
<td>Does body image differ by gender groups?</td>
<td>chi square</td>
<td>Q8, AGE, Q8 YOURTHOUGHTSONYOURBODY</td>
<td>World Health Organization, Health Behaviors of School-aged Children</td>
<td>1. Each person, item, entity is included in only one cell of the contingency table (i.e., no repeated measures). 2. Expected frequencies should be greater than 5</td>
<td>Look at row and column percentages to interpret any effects</td>
<td>Identify response rate required for sufficient power; Does sampling method address this? Time required to complete, motivation, validity of self-reported data. Prevent missing responses</td>
<td>How to handle missing and inconsistent data? Need to identify multiple types of missing data? How to handle logical errors?</td>
<td>Valid range, Expected distribution, Expected missing cases</td>
<td>Screen with frequency distributions and check standard deviation for sample and by racial/ethnic groups. Check missing cases for variables against overall missing data rate. Run normality tests (K-S or S-W). Screen with frequency distributions and check standard deviation for sample and by racial/ethnic groups. Check missing cases for variables against overall missing data rate. Run normality tests (K-S or S-W).</td>
</tr>
<tr>
<td>Is there a correlation between body image and the number of close friends?</td>
<td>scatterplot, Pearson's correlation coefficient, correlation coefficient squared</td>
<td>Q8, YOURTHOUGHTSONYOURBODY, Q8_1 NUMBER OF CLOSE FRIENDS: MALES, Q8_2 NUMBER OF CLOSE FRIENDS: FEMALES</td>
<td>World Health Organization, Health Behaviors of School-aged Children</td>
<td>1. Interval or ratio level measurement; 2. Normally distributed</td>
<td>Use Spearman's correlation coefficient if non-normal distribution</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does reported depression differ between racial/ethnic groups?</td>
<td>World Health Organization, Health Behaviors of School-aged Children</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------------------------------------</td>
<td>-------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the relationship between depression and the number of close friends?</td>
<td>World Health Organization, Health Behaviors of School-aged Children</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does the rate of bullying differ between racial/ethnic groups?</td>
<td>World Health Organization, Health Behaviors of School-aged Children</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Next time I will:

*Next Time I Will:* Not make my own wedding cake.
Bioinformatics class group exercise: The importance of reproducibility in "in silico" research

Pamela Shaw
Northwestern University
Group Exercise: The importance of reproducibility in “in silico” research
Pamela Shaw, Northwestern University Galter Health Sciences Library

Run BLASTn on a “mystery sequence”

Left side of the room
Run with defaults (megablast)
Results

Right side of the room
Change from megablast to blastn
Results
Take home message
(Group discussion)

Changing parameters for bioinformatics analysis in online platforms and software packages

Defaults change over time and with updated software versions

One change to default settings can make a big difference in results

Don’t rely on memory to recall what settings you change

Use NCBI’s options to “Save search strategies”, options in other bioinformatics sites (Galaxy, etc.) to save parameters, workflows and defaults

This enhances reproducibility = good data practice
#15
Data Management Risk Analysis
Elise Dunham
University of Illinois at Urbana-Champaign
Risk Score = Impact x Likelihood

<table>
<thead>
<tr>
<th>IMPACT</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>(Financial / Non-Financial)</td>
</tr>
<tr>
<td>5</td>
<td>Greater than $50 million Or</td>
</tr>
<tr>
<td></td>
<td>Extreme reputational impact</td>
</tr>
<tr>
<td>4</td>
<td>$25 million to $50 million Or</td>
</tr>
<tr>
<td></td>
<td>High reputational impact</td>
</tr>
<tr>
<td>3</td>
<td>$5 million to $25 million Or</td>
</tr>
<tr>
<td></td>
<td>Medium to low reputational impact</td>
</tr>
<tr>
<td>2</td>
<td>$100,000 to $5 million Or</td>
</tr>
<tr>
<td></td>
<td>Low to no reputational impact</td>
</tr>
<tr>
<td>1</td>
<td>Less than $100,000 Or</td>
</tr>
<tr>
<td></td>
<td>No reputational impact</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LIKELIHOOD</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>(Impact and Likelihood reflect existing controls)</td>
</tr>
<tr>
<td>5</td>
<td>Almost certain; expected to occur</td>
</tr>
<tr>
<td>4</td>
<td>Likely; probably will occur</td>
</tr>
<tr>
<td>3</td>
<td>Possible; might occur at some time</td>
</tr>
<tr>
<td>2</td>
<td>Unlikely; could occur at some time</td>
</tr>
<tr>
<td>1</td>
<td>Rare; may occur</td>
</tr>
</tbody>
</table>


Elise Dunham, University of Illinois at Urbana-Champaign (imker@illinois.edu)
# Immediate Response Strategies

<table>
<thead>
<tr>
<th>Risk Score *</th>
<th>Consequences</th>
<th>Immediate Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very High Risk (21-25)</td>
<td>Extreme financial loss; extreme reputational impact</td>
<td>Requires essential and immediate allocation and organization of resources to manage/mitigate the risk; establish plans and countermeasures.</td>
</tr>
<tr>
<td>High Risk (16-20)</td>
<td>High financial loss; high reputational impact</td>
<td>Requires priority allocation of resources for management and/or mitigation; establish plans and countermeasures.</td>
</tr>
<tr>
<td>Moderate Risk (11-15)</td>
<td>Moderate financial loss; medium to low reputational impact</td>
<td>Allocation of resources for study is desirable; risk should be monitored for increases in impact or likelihood.</td>
</tr>
<tr>
<td>Low Risk (6-10)</td>
<td>Low financial loss; low to no reputational impact</td>
<td>Generally does not require action, but should be reviewed periodically.</td>
</tr>
<tr>
<td>Very Low Risk (1-5)</td>
<td>Negligible financial loss; no reputational impact</td>
<td>No action required.</td>
</tr>
</tbody>
</table>

* Risk Score = Impact (1 to 5) x Likelihood (1 to 5)


Elise Dunham, University of Illinois at Urbana-Champaign (imker@illinois.edu)
#16
Peer Review of Documentation
Active Learning Classroom Activity
Megan Sapp Nelson
Purdue University Libraries
Peer Review of Documentation Classroom Activity

Preliminaries:
   Introduce yourself.
   Exchange documentation/code.
   Two minute explanatory conversation about documentation. What’s it supposed to describe/do?

Round 1 – Content:
   What design decisions are present?
   Why were the decisions made?

Discuss best practices/ communication inhibitors.

Round 2 – Format:
   What is confusing?
   What are the barriers to communication?
   What communicates well?

Discuss best practices/ communication inhibitors.

Megan Sapp Nelson, Purdue University Libraries (msn@purdue.edu)
#17
(Public) Post Publication Review
Heidi Imker
University of Illinois at Urbana-Champaign
• Dedicated Sites
  – F1000Prime
  – PubMed Commons
  – ScienceOpen
  – PubPeer
  – Publons
  – The Winnower
  – Retraction Watch
• Personal Blogs
• Twitter
Selfishness as a Motivator for Data Management

Jamene Brooks-Kieffer
University of Kansas
Developing Skills in Research Data Services

Self-directed methodology applicable to any discipline: exploration of mandates, repositories, metadata standards, other elements of disciplinary data culture

General Data Concepts

Establishing a vocabulary with basic topics:
- Formats, storage, security, sharing, DMPs

Deep Dive

In-depth topics that transcend any one discipline:
- Data reference, text mining, International data, etc.

Advanced Workshops
Adapting metadata across disciplines

Helenmary Sheridan
Northwestern University Library
Recognizing value of humanities research data

Creator: unknown Ladaki (Ladakhi cultural designation)
Title: Wanla Monastery. Exterior view: buildings on cliff with stupas in foreground.
Date: unknown
Location: India; Jammu and Kashmir; Ladakh; Wanla.
Source: Rob Linrothe. 1996.

#21

Shared Brainstorming

Kaitlin Svabek
University of Wisconsin-Madison SLIS
SHARED BRAINSTORMING

Get into small discussion groups of four or five. Have a group member jot down notes. Be ready to share your ideas with the larger group.
#22
Code is data too. Git, GitHub and Software Carpentry
Mark Laufersweiler
University of Oklahoma
Software Carpentry
http://software-carpentry.org/

• “The Software Carpentry Foundation is a non-profit volunteer organization whose members teach researchers basic software skills.”

• SWC runs over a hundred workshops a year, builds and maintains open access teaching materials, and runs an instructor training program”

• SWC as moved beyond STEM groups
  • Librarians - cataloging librarians for Python (MARC)
  • Humanists – Natural Language toolkits for Python and R

• Git and GitHub modules offered as part of 2 day workshop
  • Git is a version control system that can be used with code or text files and has been used with small data sets (not recommended)
GitHub
https://github.com

• GitHub is a Git repository hosting system
  • Web based
    • Free – no private repositories
    • Education – free and adds two private repositories
    • Group – paid and adds 10 private repositories

• Researchers can link software releases to DOI’s to create citable references
  • Zenodo (https://zenodo.org)
  • Versions of the code can be then tied to research papers published
  • Other researchers can track the providence of the code if it has been developed further since the published paper.
Bridging Researchers’ Active Data Storage Needs
Matt Schultz
Grand Valley State University Libraries
Bridging Researchers’ Active Data Storage Needs

- What type of data and data formats are you creating, analyzing & managing?
- How much data (MB/GB/TB etc.) do you anticipate amassing and over what time-frame?
- How many researchers and how many will need access to active storage? Are access roles clearly defined?
- What is your IRB protocol (Exempt/Expedited/Full)? Do you have HIPAA and/or FERPA restrictions?
- What does your funding agency require to ensure research subject protections?
- What institutional storage resources are available? What role can hosted services play if any? What are the cost gaps?
- Do the above solutions ensure adequate privacy, security & access permissions control?
- Do you use good password and encryption strategies for electronic research data? Good security strategies for analog data?

Matt Schultz, Grand Valley State University Libraries (schultzm@gvsu.edu)
More Than a Spreadsheet: Documenting the Data You Download

Kristin Partlo
Carleton College

*Don’t Forget the Metadata*
Assigned Groups and Sources to Explore

<table>
<thead>
<tr>
<th>Group Members</th>
<th>Source to Explore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1: Quasi-time series</td>
<td>Eurostat <a href="https://ec.europa.eu/eurostat/web/budget-surveys">Household Budget Surveys</a> or <a href="https://ec.europa.eu/eurostat/web/indicators">Statistics by Theme</a></td>
</tr>
<tr>
<td>Stephen, Becca, Peter</td>
<td>--------------------------------------------------------</td>
</tr>
<tr>
<td>Group 2: Finance</td>
<td>IMF <a href="https://www.imf.org/external/data.aspx">Data by Topic</a></td>
</tr>
<tr>
<td>Group 3: Commodity-dependent countries</td>
<td>David Jacks' data from his NBER working paper &quot;From Boom to Bust&quot;</td>
</tr>
<tr>
<td>Laura, Gabbi</td>
<td>--------------------------------------------------------</td>
</tr>
<tr>
<td>Group 4: Inequality panel</td>
<td>UNU-Wider Inequality Data or FDI data from World Bank</td>
</tr>
<tr>
<td>Dylan, John, Anil</td>
<td>--------------------------------------------------------</td>
</tr>
<tr>
<td>Group 5: Growth and Happiness</td>
<td>World Values Survey</td>
</tr>
<tr>
<td>Ben, Jason</td>
<td>--------------------------------------------------------</td>
</tr>
<tr>
<td>Group 6: Growth and Inequality</td>
<td>World Bank <a href="https://data.worldbank.org">Poverty &amp; Equity Database</a> or <a href="https://data.worldbank.org">World Development Indicators</a></td>
</tr>
<tr>
<td>Hayden (or join Group 4 or 5)</td>
<td>--------------------------------------------------------</td>
</tr>
</tbody>
</table>

Instructions:
Explore, then download.

3 Sets of Prompts:
- Summarize the source
- Find the documentation
- Identify the download format options

Timeline
10 minutes group work
5-10 minutes discussion

For Your Research Journal
Checklist of what to record about a data download:
- URL
- Date
- Author, principal investigator, producing agency, etc.
- Exact name of dataset, not the website's name and the version if appropriate
- What you had to query in order to get the exact download you got
- Make local copies of all relevant documentation;
- Suggested citation information

Kristin Partlo kpartlo@carleton.edu
https://goo.gl/HksYSA
#25
Data Modules (Pilot) Project
Aaron Albertson / Ron Joslin
Macalester College
Data Modules (Pilot) Project

- **Online, self-paced modules and activities** covering such topics as metadata creation, documentation, backing-up/security, sharing/distribution, publication, file organization, naming conventions, finding/selecting data, copyright, and access.

- **Intended to coincide with data research project** assignment in class; students complete modules as they are completing assignment, get feedback from librarians outside; options for having one-on-one 'data consultations' with a librarian or having a librarian conduct an in-class instruction session also offered.

- **Modules 'chunked'** to allow faculty to select the specific topics their students complete.

- **Piloting with two classes this semester**: Statistical Modeling class and Political Science Empirical Research class.

Much of the content in the modules has been adapted from the following sources:

Confused? Concept map it!

Carrie Leatherman
Western Michigan University
#29
Discussion Activity for Identifying Data Risks

Josh Bishoff
University of Minnesota
Documentation exercise

This project is going to gather images of wildlife in the Superior National Forest throughout the month of November. A series of 35 motion-activated digital cameras will be attached to trees in a variety of locations in the forest. Each week, 5 graduate students will visit 7 cameras each to check for damage and to swap out each camera’s micro SD card with a fresh, blank micro SD card. The grad students will then return the cards to the lab & transfer image files from the cards to a lab computer for further analysis.

What could go wrong?
Documentation exercise

10 minutes (small groups): identify as many potential risks to your data as you can

10 minutes (small groups): choose a particular risk, and think of a plan that addresses & reduces that risk

10 minutes (entire class): share out & discuss plans. What additional work will need to be done?

Josh Bishoff, Minnesota (jbishoff@umn.edu)
#30
Students’ Critical Reflection Form about their Big Data Problem

Line Pouchard
Purdue University Libraries
Think while you fill (out the form)

• What is the total file size, the number of files that you expect, and where will you store it?

• How secure is it?

• What are some of the formats and software involved?

• What kind of documentation will you need and produce?

• Who owns this data?

• What will you preserve?
Beekeeper Eric @BeekeeperEric

We analyzed data from the final two years of our experiment, as results from the first year are really messed up
#overlyhonestmethods

11:06 AM - 8 Jan 2013

Anne Osterrieder @AnneOsterrieder

We don't know how this method was performed because the PhD student's lab book is written in a foreign language.
#overlyhonestmethods

7:33 AM - 8 Jan 2013

Morgan Edwards @mangoedwards

I can't send you the original data because I don't remember what my excel file names mean anymore #overlyhonestmethods

12:11 PM - 8 Jan 2013
Good practices for organizing data files

• Define the types of data and file formats for the research.

• Choose a meaningful directory hierarchy/naming convention.

• Include important contextual information.

• Organize folders by primary, secondary, tertiary subject or collection method.

• Choose a naming convention and ensure that the rules are followed systematically by always including the same information in the same order.

• Document your system and use it consistently.

#32
think-pair-share!

Cameron Cook - UW-Madison
What do we have? What do we want to get? How are we going to get there?

Dan Able
Emporia State (Student)
What Do We Have?
• Given information, the “inputs”

What Do We Want To Get?
• Solution to the problem, the “outputs”

How Are We Going To Get There?
• Techniques used to solve the problem, which are influenced by the answers to the previous two questions
A local train leaves Milwaukee at noon, traveling at a speed of 40 mph. At 1:00 pm, the express leaves the same station on the same route, traveling at a speed of 60 mph. At what time will the express overtake the local?

**What Do We Have?**
The speeds of each train
The fact that the express has traveled for 1 hour less than the local
Both trains have gone the same distance at the time we’re interested in

**What Do We Want To Get?**
The time taken for the express to overtake the local

**How Are We Going To Get There?**
Distance=Rate x Time
Let x=Travel time for the local train, x-1=Travel time for the express
Distances are equal, so write equation 40x=60(x-1)
Solve: x=3 hours
So, the express overtakes the local at 3:00 pm.
Blended Learning: Online: Pre-discussion, Inclass: Ice breaker, Lecture, Small Group, Re-group

Mary Murphy
UW-Madison (Student)
#35
ORCID for Personal Data Management
Trisha Adamus
University of Wisconsin-Madison
Data Management Plans (DOE)

Laniece Miller
Argonne National Laboratory
Data Management Planning

Laniece Miller (Argonne National Laboratory)

Focused on DOE Office of Science DMP requirement

- Digital Research Data
- Size / number of files
- Source
- Validating?
- How / where Shared?
- When shared?
- Metadata describing data
- Required Software

- Sharing Responsibility
- Sharing Resources needed?
- How / where Preserved?
- Duration Preserved?
- Preservation Responsibility
- Preservation Resources needed?

Walk through table and template so they leave with notes to complete DMP

Many table headers from University of Columbia