Research Data Lightning Talks

Research Data Service Interest Group
January 27, 2016
Welcome!
Data Practices and Perspectives of Atmospheric and Engineering Faculty

CHRISTIE WILEY
ENGINEERING RESEARCH DATA SERVICES LIBRARIAN
GRAINGER ENGINEERING LIBRARY
1-26-16
Purpose of Study

What does data management look like in the context of faculty research?

The study will also explore the use of these services, identify needs and gaps in application, and determine other areas where librarians can meet data management needs.
Engineering and Atmospheric Faculty

STUDY PARTICIPANTS

INTERVIEW QUESTIONS

Current research projects
Funding sources
Type of data
Who is interested in data and how is it used
Disciplinary repository
Data sharing
Finding data
Awareness of current literature
Awareness of data management services/campus repository
Agency Review panel experiences
Struggles with data
Themes of the study

Awareness of data management preservation service
Uses of research
Campus /Disciplinary repository
Data sharing perspectives
Data documentation
Next steps

Successes

Challenges

#No Data Problems
MALLLET and Topic Modeling

- MALLET is a statistical natural language processing application that runs on Windows and Mac. Topic Modeling is using algorithms to uncover hidden thematic structure ("topics") in texts. MALLET is one of many tools that performs topic modeling.
I've had many shocking encounters with library staff. When I asked if the library would consider allowing patrons to return materials around campus, I was told that kittens are routinely killed in such return boxes... ... The library staff seems to me to share a strange perspective which is all about 1) keeping materials from users, and then 2) getting those materials back in the library ASAP. I've been disappointed with the library...
DATASETS

that I tested and learned ...

LibQUAL+ Lite survey comments: MALTET trained 10 topic models, however the MALTET models are more confusing than human coding.

Library Job Descriptions: MALTET trained 20 topic models; helpful on grouping JDs by similar job responsibilities and/or JD formats.
WHAT'S THE POINT?

if you are interested ...

- Good for filtering out stopwords and "meaningless counts". A fast and objective way of identifying potential topics in texts.
- Be mindful of the difference between "relevance" and "probability".
to get you started

- Mallet website http://mallet.cs.umass.edu
The potential of Scholarly Perspective for Managing Research Data

Pompilia Burcica, PhD
GSLIS, MS student
LLL Library, GA
• We examine government data from a citizen’s perspective - Can a citizen analyze a city budget dataset? (no...)

• We examine demographic data from a marketing perspective. Can a customer predict how data brokers target him/her based on the personal traumas like divorce, death of child? (no...)

Should such datasets be provided with a helpful explanation, interpretation, analysis? With a scholarly perspective?
A scholarly perspective is not necessary:

- When the release of Medicare payments data exposes profit-driven doctors who overuse it.
- When a dataset of the New Yorker cartoons exposes racial and gender profiling.

But:

To be understood, certain datasets need research analysis, research conclusions, need a scholarly perspective.
Then, why do research data lack a scholarly perspective?

- Upon depositing a dataset, is the scholarly perspective visible in the metadata template?
- Are other metadata descriptors based on the scholarly perspective in place?
- Did scholars depositing the research dataset frame it according to their scholarly argument?

The scholars directly should be responsible for the scholarly perspective of their dataset and not the librarians - *disintermediation*
Q: What is the scholarly output? Is it the book? Is it the article? Is it the digital collection?
A: It is the scholarly perspective.

To determine it:
• Where is the work situated in the field?
• What is the argument of the article, book, dataset?
• On what did the argument build or what did the argument challenge?
• What is the interdisciplinary foundation of the argument?
With a clear scholarly perspective, scholars can:

- manage their data more purposefully, with the argument/perspective being placed front and center
- enhance the access of others to their datasets if accompanied by a scholarly perspective
- Have their curation work benefit from metadata descriptors like: research origin descriptors, data set structural descriptors all drawing on the scholarly perspective
A Case-Study from Atmospheric Sciences

*We have:*

A dataset recording flight observations of cloud seeding

*We need:*

- A scholarly perspective – i.e. information accompanying the dataset about the theory being demonstrated, conclusions being reached, other AS subfields that might be interested in, using this dataset

*This might be found in the article abstract but it has to follow the dataset separate from articles in case of re-purposing research data.*
<table>
<thead>
<tr>
<th>Time (UT)</th>
<th>Air_Temp</th>
<th>TAS</th>
<th>POS_Pitch</th>
<th>POS_Head</th>
<th>POS_Lat</th>
<th>POS_Lon</th>
<th>POS_Alt</th>
<th>POS_Spd</th>
<th>2-DC_Conc</th>
<th>2-DC_LWC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20000</td>
<td>10000</td>
<td>100000000</td>
<td>10000000</td>
<td>1000000</td>
<td>1000000</td>
<td>1000000</td>
<td>1000000</td>
<td>1000000</td>
<td>1000000</td>
</tr>
<tr>
<td>2</td>
<td>30000</td>
<td>20000</td>
<td>200000000</td>
<td>20000000</td>
<td>2000000</td>
<td>2000000</td>
<td>2000000</td>
<td>2000000</td>
<td>2000000</td>
<td>2000000</td>
</tr>
</tbody>
</table>

**Notes:**
- **Time (UT):** UT seconds from midnight on day aircraft flight started (based on Data System Time).
- **Air_Temp:** Air Temperature (Dewpoint + 5°C) [°C].
- **TAS:** True Air Speed (Blended First on the main temperature instrument and secondarily on the backup temperature instrument) [m/s].
- **POS_Pitch:** Aircraft pitching angle from the Aircraft Position and Orientation System (POS) [degrees].
- **POS_Head:** Aircraft heading angle from the Aircraft Position and Orientation System (POS) [degrees].
- **POS_Lat:** Aircraft latitude from the Aircraft Position and Orientation System (POS) [degrees].
- **POS_Lon:** Aircraft longitude from the Aircraft Position and Orientation System (POS) [degrees].
- **POS_Alt:** Aircraft altitude from the Aircraft Position and Orientation System (POS) [m].
- **POS_Spd:** Aircraft ground speed from the Aircraft Position and Orientation System (POS) [m/s].

**Additional Data:**
- **2-DC_Conc:** Number concentration of droplets based on the 2-DC Probe measurements [m⁻³].
- **2-DC_LWC:** Liquid Water Content based on the 2-DC Probe measurement including cloud liquid water content (g/m³).
- **2-DC_Temperature:** Temperature from the 2-DC Probe [°C].
- **2-DC_Velocity:** Velocity from the 2-DC Probe [m/s].

**Final Data:**
- **Time:** Time at which the data was collected.
- **Air_Temp:** Air Temperature (Dewpoint + 5°C) [°C].
- **TAS:** True Air Speed (Blended First on the main temperature instrument and secondarily on the backup temperature instrument) [m/s].
Data Citation Case Study – National Atmospheric Deposition Program

Susan Braxton
A Data Citation Case Study

• NADP began in 1977 as a project of US State Agricultural Experiment Stations

• Goal: “providing data on the amounts, trends, and geographic distributions of acids, nutrients, and base cations in precipitation”

• Currently 5 networks measuring parameters including acidity, conductance, calcium, magnesium, sodium, potassium, sulfate, nitrate, chloride, ammonia, ammonium, and mercury

• Work is broadly collaborative and funded by multiple agencies, institutions, organizations

• Data are freely available online

Braxton, RDSiG 1/27/2016
Finding where the data are used

• Some traditional products (e.g., annual maps) are cited traditionally.
• Some researchers have registered data subsets retrieved/used with DataCite, although it is not clear they cited them traditionally.
• Often, NADP is mentioned in table/figure captions and not referenced in the bibliography.
• Google alerts retrieve works that mention NADP.
• Manual review of each work is done to verify NADP data use.
• 2007-2014, gathered 1,277 works all confirmed to have used NAPD data (includes articles, books/chapters, proceedings papers, reports, and dissertations/theses)
Visualization tools

- **VOSviewer**
  - Bibliographic visualization tool
  - Free application download or web instance.
  - Author, institution, journal, topic analysis.

- **Pushgraph**
  - From ChalkLabs, Bloomington, IN.
  - *Not* free.
  - We used National Institute of Food and Agriculture (NIFA, USDA) custom subject concept map based on 130,000 NIFA project documents.
Second Round of Retrieval...

• Both tools expected spreadsheet / csv formatted metadata, and richer metadata than we had from Google Scholar
• We chose Scopus to retrieve metadata for analysis
  • Broad subject coverage
  • Output format compatible with tools
  • Allowed saving and sharing lists from session to session
  • We can augment later from other sources
• Some consequences...
  • Bias toward articles
  • Bias against niche journals?
Some Preliminary Results...

• 963 of the known 1,277 works were found in Scopus
• Citation of NADP data in bibliographies <50% for Scopus set
Visualization examples...

USDA NIFA Pushgraph™ concept map (grey) with NADP papers (colored dots) overlaid. Dot color indicates frequency.
For additional information...

• Knighton, R., S. Braxton, D.A. Gay, and N.S. Nasarudin. Visualization of Science Supported by NADP Measurements [poster]. Acid Rain 2015, 9th International Conference on Acid Deposition, 19-23 October 2015, Rochester NY. http://hdl.handle.net/2142/88854

• National Atmospheric Deposition Program http://nadp.isws.illinois.edu/

• VOSviewer, now on version 1.6.3 http://www.vosviewer.com/

• ChalkLabs Pushgraph overview http://chalklabs.com/?page_id=865
(Public) Post-Publication Review

Heidi Imker
Director, Research Data Service
(Public) Post Publication Review

- Dedicated Sites
  - F1000Prime
  - PubMed Commons
  - ScienceOpen
  - PubPeer
  - Publons
  - The Winnower
  - Retraction Watch
- Personal Blogs
- Twitter
The Irish Potato Famine Pathogen Phytophthora infestans Translocates the CRN8 Kinase into Host Plant Cells
van Damme et al
PLoS Pathogens 2012

Posted on PubPeer
Take-Aways

1. New expectations for greater transparency with associated technologies, platforms, and policies for enabling it.
2. Fraction targeted likely will be small, but be prepared for...
RDSIG Save the Date: Data Conference Reports

Wednesday, February 24, 2016
3pm-4pm
Thank you!