

Research Data Lightning Talks

Research Data Service Interest Group
January 27, 2016

RESEARCH
DATA SERVICE

at

UNIVERSITY LIBRARY
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN



Welcome!

Data Practices and Perspectives of Atmospheric and Engineering Faculty

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



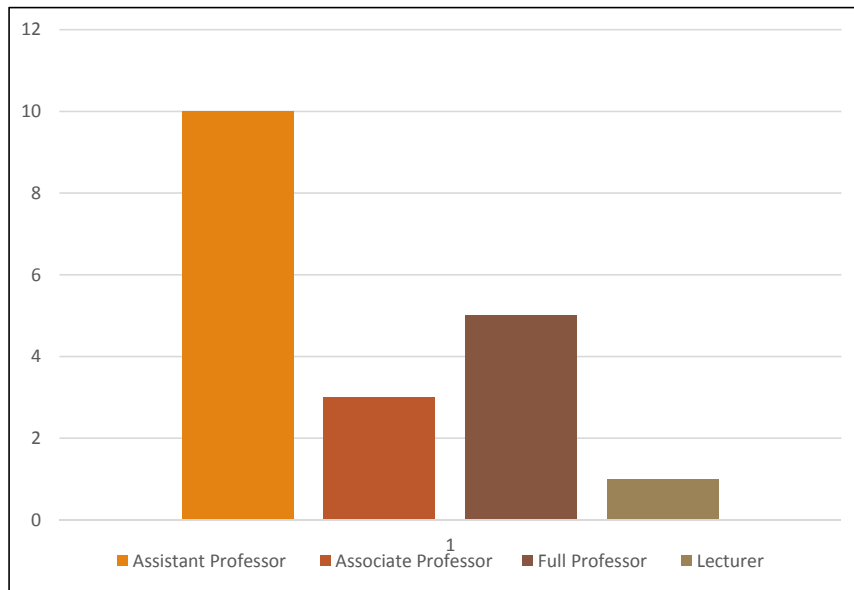
Purpose of Study

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

FUN WITH

MALLET *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.

JEN-CHIEN YU

PROBLEM

Coding survey comments

\\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 ...

JEN-CHIEN YU

DATASETS

that I tested and learned ...

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

\\Library Job Descriptions: MALLET 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".

RESOURCES

to get you started

- MALLET website <http://mallet.cs.umass.edu>
- Getting Started with Topic Modeling and MALLET <http://programminghistorian.org/lessons/topic-modeling-and-mallet>
- Topic Modeling <https://www.cs.princeton.edu/~blei/topicmodeling.html>

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.

Cap File Citation Return Flight Grand Forks Dataset.xlsx - Microsoft Excel

File Home Insert Page Layout Formulas Data Review View

Clipboard Font Alignment Number Styles Cells

Calibri 11 A⁺ A⁻ B I U Merge & Center General Conditional Formatting Format as Table Cell Styles Insert Delete Format

A38	Time	Air_Temp	TAS	POS_Pitch	POS_Head	POS_Lat	POS_Lon	POS_Alt	POS_Spd	2-DC_Conc	2-DC_LWC
2	Delene, David										
3	University of North Dakota										
4	UND Citation II (N556DS)										
5	UTC2014										
6	11										
7	2015 08 11 2015 11 01										
8	1										
9	Time [seconds]; UT seconds from midnight on day aircraft flight started (Based on Data System Time)										
10	21										
11	1.0000	1.0000	1.00000000	1.00000000	1.00000000	1.000000	1.000000	1.0000	1.0000	1.0000	1.0000
12	999999.9999	999999.9999	999.9999999	999.9999999	999.9999999	9999.99999	99999.99999	99999.99999	99999.99999	99999.99999	999999.9999
13	Air Temperature Corrected for Dynamic Heating (Based first on the main temperature/pitot instrument and secondarily based on the backup temperature/pitot instrument) [degC]										
14	True Air Speed (Based first on the main temperature/pitot instrument and secondarily based on the backup temperature/pitot instrument) [m/s]										
15	Aircraft pitch angle from the Applanix Position and Orientation System (POS); -180 to 180 range with 0 being level and positive angles in the clockwise (upward) direction away from center of the Earth [degrees]										
16	Aircraft heading angle from the Applanix Position and Orientation System (POS); 0 to 360 range with 0 being North and angles increasing in a clockwise (right) direction [degrees]										
17	Aircraft latitude from the Applanix Position and Orientation System (POS); -90 to 90 range with positive values in Northern Hemisphere and negative values in Southern Hemisphere [degrees]										
18	Aircraft longitude from the Applanix Position and Orientation System (POS); -180 to 180 range with positive values in Eastern Hemisphere and negative values in Western Hemisphere [degrees]										
19	Aircraft altitude from the Applanix Position and Orientation System (POS) [m]										
20	Aircraft ground speed from the Applanix Position and Orientation System (POS) [m/s]										
21	Number concentration of droplets based on the 2-DC Probe measurements [# /cm^3]										
22	Equivalent liquid water content based on the 2-DC Probe measurements assuming all liquid hydrometeor [g/m^3]										
23	Number concentration of droplets based on the 2-DC Probe measurements for hydrometeors greater than 105 um [# /cm^3]										
24	Equivalent liquid water content based on the 2-DC Probe measurements assuming all liquid hydrometeor for hydrometeors greater than 105 um [g/m^3]										
25	Total Water Content based on the Nevzorov Probe measurement										
26	Liquid Water Content based on the Nevzorov Probe measurement without correction for residual ice										
27	Liquid Water Content based on the Nevzorov Probe measurement with correction for residual ice (beta = 0.110000) [g/m^3]										
28	Ice Water Content based on the Nevzorov Probe measurement (beta = 0.110000) [g/m^3]										
29	Dewpoint Temperature from EG&G Probe [degC] {Calibration: slope = 20.000000 offset = -70.000000}										
30	Dew Point Temperature from the Laser Hygrometer [degrees Celsius]										
31	Frost Point Temperature from the Laser Hygrometer [degrees Celsius]										
32	Relative Humidity from the Laser Hygrometer [percent]; With Respect to water T >= 0 Respect to Ice T < 0										
33	Relative Humidity from the Laser Hygrometer [percent]; With Respect to water even below freezing										
34	0										
35	4										
36	Final Data (Average is missing value code only if all values within the time period are missing value code.)										
37	1Hz Data										

38	Time	Air Temp	TAS	POS_Pitch	POS_Head	POS_Lat	POS_Lon	POS_Alt	POS_Spd	2-DC_Conc	2-DC_LWC	2-DC_Conch	2-DC_LwCh	New TwC	New LwC	New LwCcor	New IvC	DEWPT	DewPoint	FrostPoint	RH	RH w					
39	seconds	degC	m/s	degrees	degrees	degrees	degrees	m	MPsec	#/cm^3	g/m^3	#/cm^3	g/m^3	g/m^3	g/m^3	g/m^3	g/m^3	degC	C	C	percent	percent					
40	86002.6000	25.4594	12.2120	999.9999999	999.9999999	999.9999999	9999.99999	99999.99999	99999.99999	0.000000	0.000000	0.000000	0.000000	0.000000	999999.9999	999999.9999	999999.9999	999999.9999	999999.9999	999999.9999	999999.9999	999999.9999	11.7937	11.7937	42.7916	42.7916	
41	86003.6000	25.4627	12.2120	999.9999999	999.9999999	999.9999999	9999.99999	99999.99999	99999.99999	0.000000	0.000000	0.000000	0.000000	0.000000	999999.9999	999999.9999	999999.9999	999999.9999	999999.9999	999999.9999	999999.9999	999999.9999	999999.9999	11.8357	11.8357	42.4472	42.4472



Data Citation Case Study – National Atmospheric Deposition Program

Susan Braxton

RESEARCH
DATA SERVICE

at

UNIVERSITY LIBRARY
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

A Data Citation Case Study



National Atmospheric Deposition Program

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


**Deadly Rain Imperils
2 Adirondacks Species**

Acidic rain threatens aquatic ecosystems



Acid-tainted rain could harm Illinois' soybean crop

- 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

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.
-  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 NADP data (includes articles, books/chapters, proceedings papers, reports, and dissertations/theses)

Visualization tools

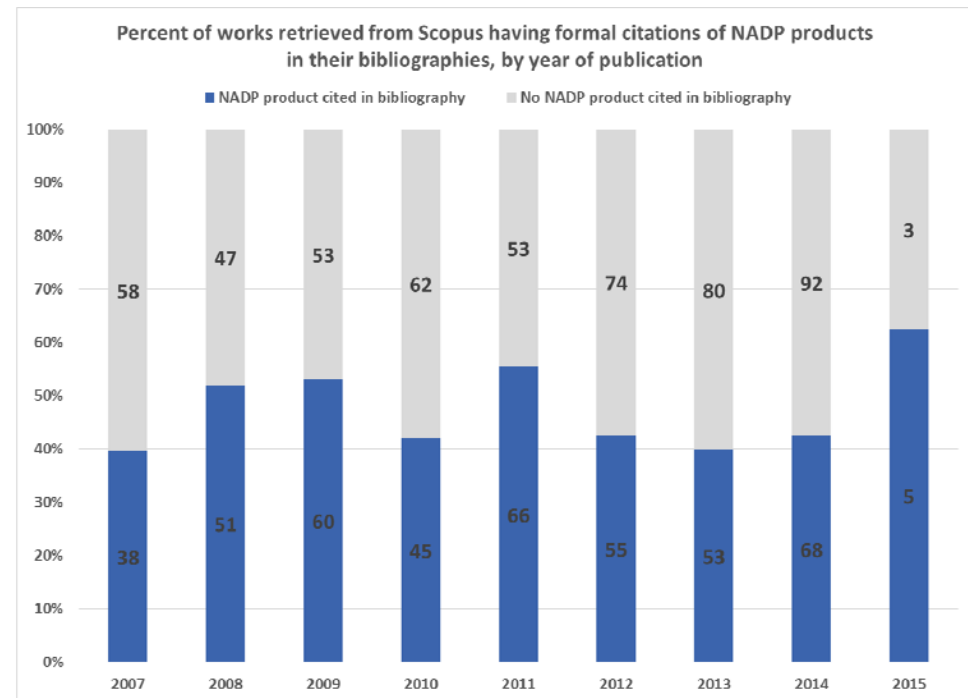
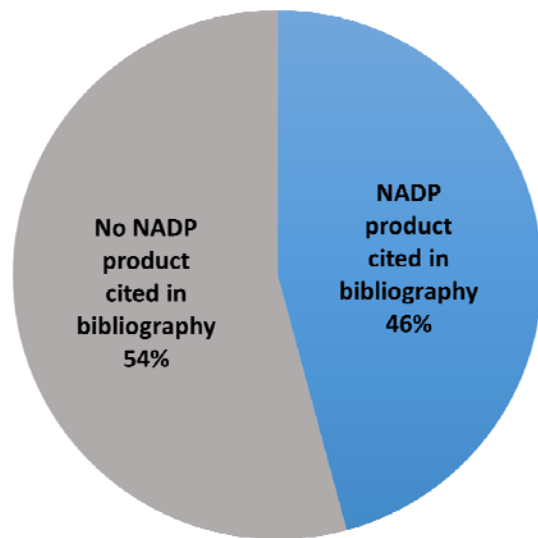
-  Bibliographic visualization tool
Free application download or web instance.
Author, institution, journal, topic analysis.
Van Eck, N. J., & Waltman, L. (2015). VOSviewer Version 1.6.2. Center for Science and Technology Studies, Leiden University. <http://www.vosviewer.com>
-  From ChalkLabs, Bloomington, 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...

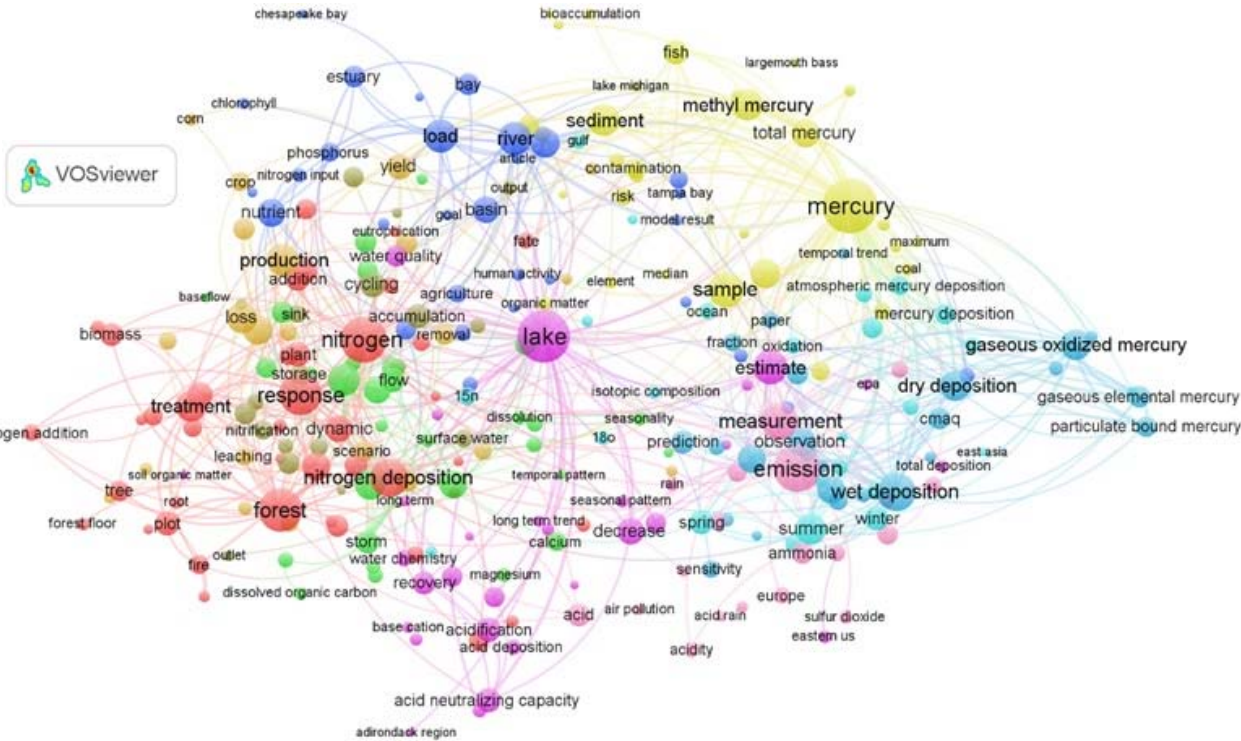
- Both tools expected spreadsheet / csv formatted metadata, and richer metadata than we had from Google Scholar
- We chose Scopus to retrieve metadata for analysis
 - Broad subject coverage
 - Output format compatible with tools
 - Allowed saving and sharing lists from session to session
 - We can augment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

Some Preliminary Results...

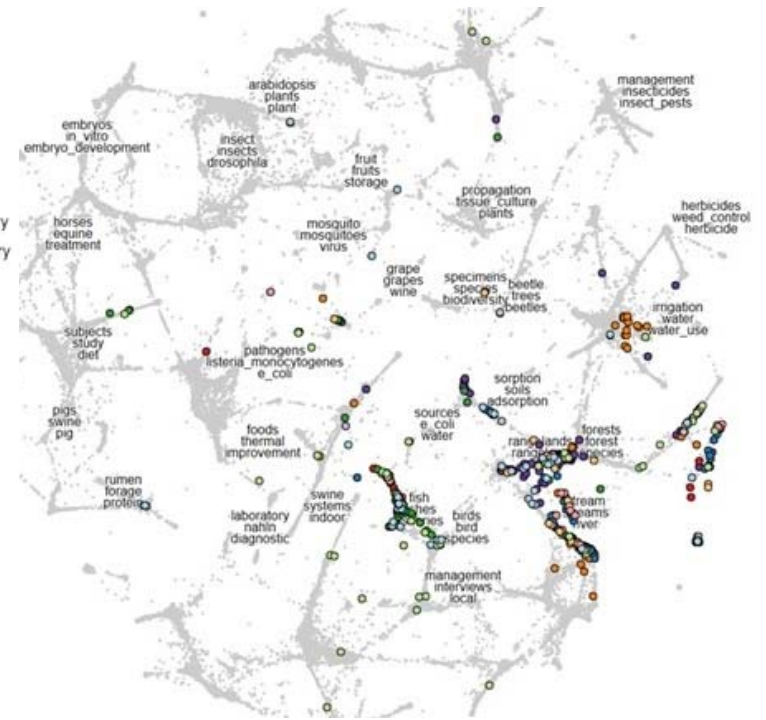
- 963 of the known 1,277 works were found in Scopus
- Citation of NADP data in bibliographies <50% for Scopus set



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

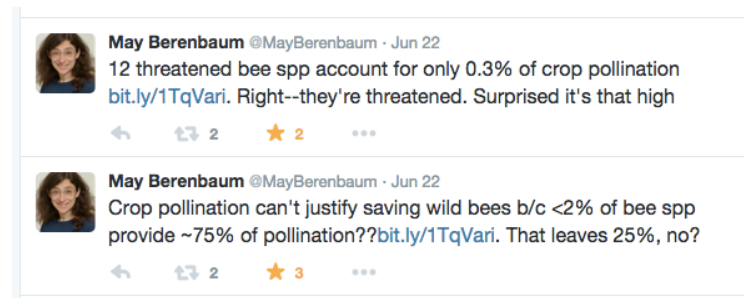
RESEARCH
DATA SERVICE

at

UNIVERSITY LIBRARY
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

(Public) Post Publication Review

- Dedicated Sites
 - F1000Prime
 - PubMed Commons
 - ScienceOpen
 - PubPeer
 - Publons
 - The Winnower
 - Retraction Watch
- Personal Blogs
- Twitter





Post Publication Review via Commentary

**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**
<https://pubpeer.com/publications/D3DF6B4B19C383804BB3748B60604D>

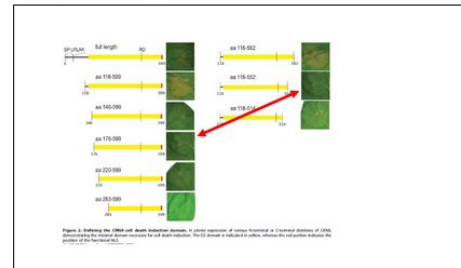
"The Irish Potato Famine Pathogen *Phytophthora infestans* Translocates the CRN8 Kinase into Host Plant Cells"

Mireille van Damme, Tolga O. Bozkurt, Cahid Cakir, Sebastian Schornack, Jan Sklenar, Alexandra M. E. Jones, Sophien Kamoun, PLoS Pathogens (2012)

Comments (19):

0

Unregistered Submission: (January 18th, 2015 5:58am UTC)



I would like to aware the authors that one of the figure panels in Figure 2 seems to be used twice in the same figure. Probably an mistake?
<http://i.imgur.com/AXkccDC.jpg>

Reply

Report

Permalink

Peer 1: (January 18th, 2015 8:14am UTC)

Seems to me that these photos are similar but not identical (e.g., in the thickness of the vein that runs diagonally from bottom right.)

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...



[Niabot](#) CC BY-SA 3.0



RDSIG Save the Date: Data Conference Reports

Wednesday, February 24, 2016
3pm-4pm



MIDWEST DATA LIBRARIAN SYMPOSIUM

hosted by UW-Milwaukee + UW-Madison

OR2015





Thank
you!