LIBER Webinar: Data Curation from a Practical Perspective

HOSTS

Dr Birgit Schmidt

Head of Knowledge Commons
Göttingen State and University
Library

LIBER Executive Board Member

bschmidt@sub.uni-goettingen.de

Rob Grim

Economics (Data) Librarian
Erasmus University Rotterdam

rob.grim@eur.nl

SPEAKER

Juliane Schneider

Team Lead/Lead Data Curator
Harvard Catalyst

Juliane_Schneider@hms.harvard.edu

NOTES

- **The webinar is being recorded.**

- **Slides and a recording will be shared** by email after the webinar.

- **Questions?** Put them in the chat box.

- **10-15 minutes of discussion** will take place following the presentations.

REAL STORIES, REAL DATA

MY TIME IN RESEARCH DATA MANAGEMENT

Juliane Schneider
Team Lead/Lead Data Curator
www.eagle-i.net
Harvard Catalyst | Clinical and Translational Science Center
Juliane_Schneider@hms.harvard.edu
https://orcid.org/0000-0002-7664-3331

What You Need to Know

- My experience is all in the United States

- I'm a metadata/discovery expert, so my stories will be skewed towards that side of data curation

- But I'll try to weave in my observations from the other types of data services
  - Meeting funding requirements
  - Finding the correct repository for data

Six Years Ago

- Institutional repositories were on the rise

- ISO-19115 (https://www.iso.org/standard/67039.html) has not yet widely replaced FGDC (https://www.fgdc.gov/standards/projects/framework-data-standard/framework-data-standard) for geospatial metadata standards required by funders

- FAIR Data Principles were not yet published

- Privacy standards were not in the conversation apart from HIPAA (Health Insurance Portability and Accountability Act of 1996)

- Data science training was just getting started as something that is crucial to good research data management

So I Hear You Want a Repository

RRZE Icon Set (CC: BY-SA)
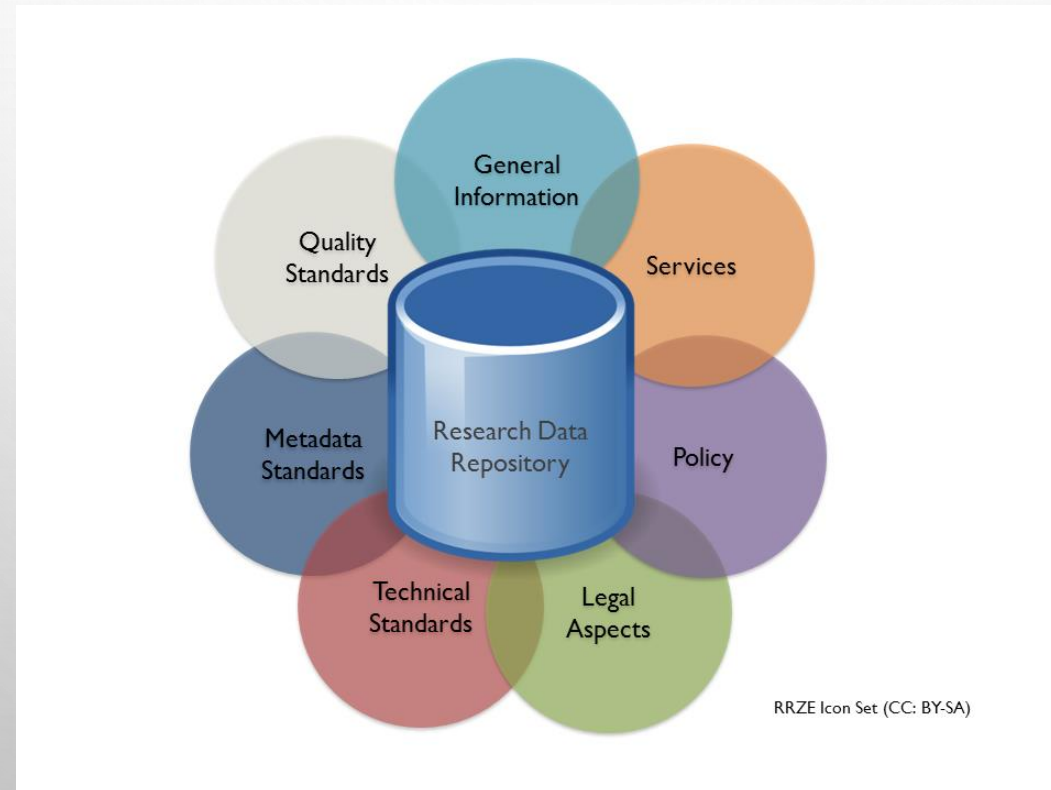
Open Platform vs. Proprietary System

- Flexible
- Change quickly to meet new requirements
- Ostensibly cheaper
- Can have a voice in the community for direction of growth
- The data within the system is owned and available to the institution

- Resource-heavy
- If you are smaller institution, your voice in the community may not be effective
- Updates and bugs can stall the repository for months
- If the community moves on, you no longer have that support
- As tech people leave, the knowledge about the system goes with them

- Stable
- Support and troubleshooting available
- User group support

- Expensive
- Not as flexible
- Data may be stuck in the system
- May not be interoperable with other systems researchers in your institution use

What Are You?

- Depository
  - Fast, minimal metadata
  - Used for funding/publication requirements
  - Research in progress
- Archive
  - Extensive metadata
  - For emeritus collections
  - Research that is finished

Why not both?

- Have minimal required metadata for immediate data sharing needs

- Offer detailed data models for archival collections

Selling vs. Listening

LOSE YOUR ASSUMPTIONS

We Assumed Researched Wanted:

A repository
Permanent Identifiers
Detailed Metadata
Complex Data Models
Data Security and Protection

Researchers Wanted:

Help with standards
Help understanding funding requirements
Help moving large sets of data around
Storage
Storage
Storage
Storage

Roadblocks

- A 3,000 acre campus plus Scripps Institute of Oceanography
- Data was stored on un-networked servers
- There was no way to securely transfer data
- Many researchers did not use the administrative email they were given, and were using the Google suite of products
- No university-wide mandate regarding data
- No researcher profiles system

Language!

ARK (Archival Resource Key)

Also Ark: Quest Object in Blockbuster Movie with Archeologist Hero

Who Are Your Partners?
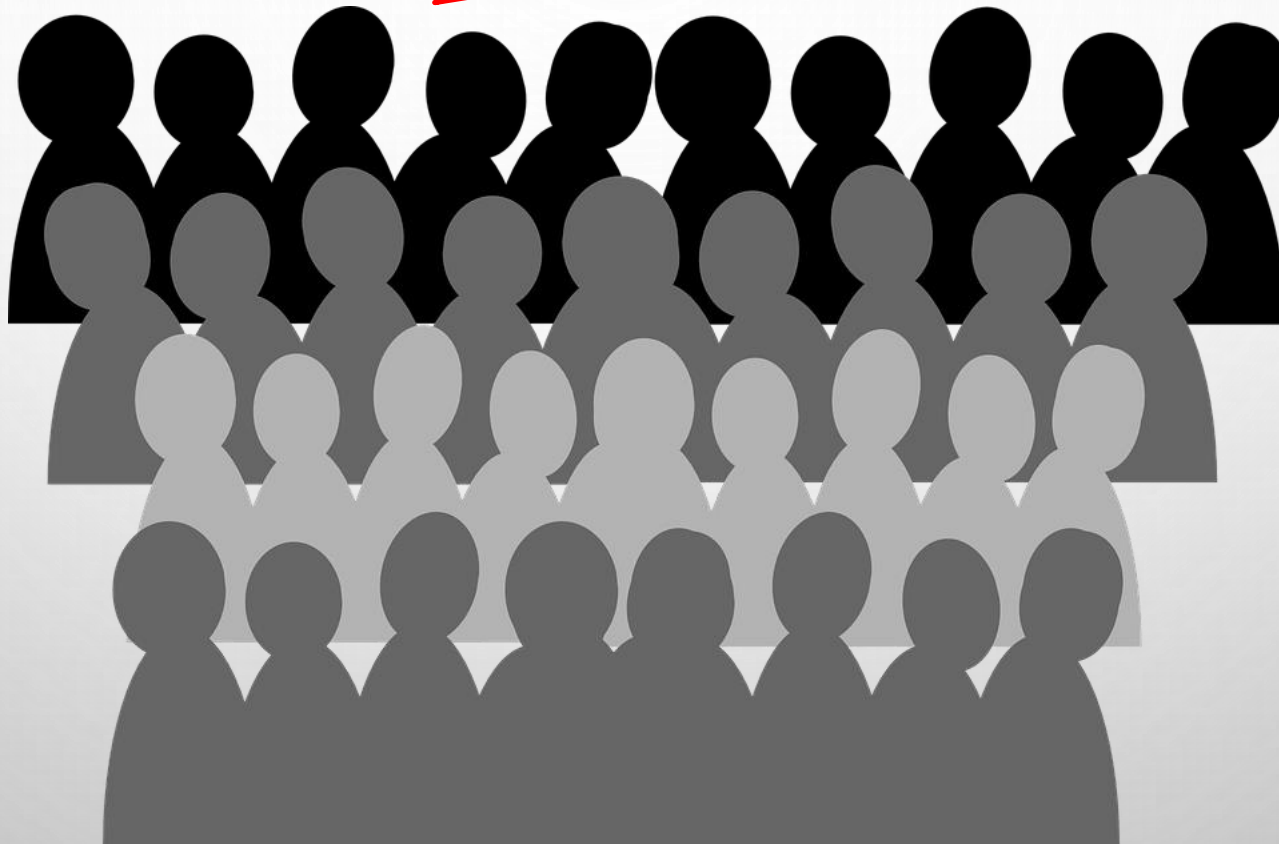
- Libraries
- Funding office
- Office of scholarly communication
- Research computing
- Core facilities
- Student groups

What Is Success?

(be realistic and kind to yourselves)

ONR Cd-79 data   (from: UCLA)

CONCENTRATION   EASTROPAC HALOBATES SAMPLES   L. CHENG 5-11-79   DATE RUN   06/22/79
IN DRY TISSUE   VALUES ARE PERCENT (%) OR PARTS PER MILLION   (1%= 10,000 PPM)   DK2:051179.C13:1

SAMPLE	WEIGHT (Dry mg)	1 P	2 NA	3 K	4 CA	5 MG	6 ZN	7 CU	8 FE	9 MN	10 B	11 AL	12 SI	13 TI	14 V	15 CO	16 NI
001MI17N119W	4.5	2812.9	649.3	**	4618.3	4218.7	382.8	104.4	215.7	28.2	26.8	107.1	262.1	8.7		**	6.2
002MI06N112W	11.5	3331.5	764.2	163.8	4931.2	3611.7	291.7	87.0	220.3	4.5	54.2	32.7	73.1	5.4	2.6		2.3
003MI05N119W	7.7	6031.5	537.4	**	8962.2	4112.0	84.2	103.1	280.5	11.7	36.3	59.6	284.4	5.3	**		
004MI13N119W	8.4	3037.5	313.4	184.0	3436.6	2900.4	93.2	72.2	234.4	8.6	35.1	106.6	164.0	2.5	1.2		1.8
005MI09N119W	7.4	2703.6	203.7	**	1674.1	2552.5	76.9	78.7	197.0	7.3	16.8	59.7	60.9				1.5
006MI02N112W	8.5	3274.9	471.3	191.8	1804.6	2990.9		64.0	405.0	10.4	22.5	187.2	92.5	**	1.7		**
007MI09N105W	8.6	5930.1	634.8	276.1	5776.9	4738.2	49.2	42.6	341.1	17.0	57.1	39.2	81.1	3.1	2.0		**
008MI04N112W	11.6	3252.8	420.4	472.5	7784.8	3630.9	93.0	49.1	302.1	6.2	67.9	27.9	113.9	4.0	1.5		1.9
009MI07N112W	11.7	2930.5	297.9	**	4668.0	2074.1		37.1	114.4	4.6	33.1	13.6	37.8	0.6			
010MI10N112W	12.5	7563.9	363.7	395.9	8.81%	6375.6	127.9	98.3	205.7	8.0	50.6	24.9	54.3	2.9	**	1.6	6.1
011MI10N119W	9.9	3000.7	777.7	**	4003.1	3606.4	121.4	58.0	172.8	9.0	44.0	37.2	55.9	0.6	**		
012MI11N112W	10.3	4297.9	801.8	263.6	2154.3	3529.6		64.5	316.3	8.5	24.2	77.1	48.1				
013MI06N119W	15.1	2416.6	751.9	**	3486.0	2313.7	104.9	61.1	123.4	2.3	31.9	25.4	33.5	0.9	**		0.5
13AMI06N119W	11.7	2501.7	940.3	**	4923.6	3030.8		67.9	166.1	3.0	58.0	39.0	88.4	1.9			
014MI07N119W	9.8	2647.0	400.4	306.3	5020.1	2831.4	183.3	48.8	172.4	6.7	12.8	68.3	59.9		**		1.3
015MI08N126W	10.7	4.73%		1013.6	14.22%	6083.5	293.1	33.9	144.7	33.3	42.3	113.5	231.7	2.7	2.8	2.1	19.3
016MI03S112W	8.7	4478.6	274.5	207.6	2566.5	2207.8	75.5	31.5	308.2	6.1	19.6	48.9	63.6				
017MI07N119W	7.9	2507.4	435.1	**	5791.8	4747.0	158.2		423.9	13.2	63.8	37.3	130.0	3.9	1.4	**	11.7

```
sequence,specimen_ID,species,latitude,longitude,mean_conc,metal
34,033,SE,20N,-119W,7.721,Cr
35,034,SE,-19S,-105W,4.439,Cr
36,035,SE,20N,-119W,7.044,Cr
37,036,SE,-13S,-126W,3.049,Cr
38,037,SE,-16S,-126W,53.431,Cr
39,038,SE,-15S,-126W,13.526,Cr
40,039,SE,-17S,-126W,9.722,Cr
41,040,SE,-14S,-126W,7.440,Cr
42,041,SE,-20S,-119W,6.919,Cr
43,042,SE,-20S,-125W,13.803,Cr
44,043,SE,-20S,-126W,9.575,Cr
45,044,SE,20N,-126W,13.384,Cr
46,045,SE,-15S,-119W,10.561,Cr
47,046,SE,-15S,-119W,3.930,Cr
48,047,SE,-18S,-119W,7.876,Cr
49,048,SE,-20S,-122W,8.268,Cr
123,120,SE,12N,-145W,1.510,Cr
159,157,SE,21N,-158W,0.687,Cr
160,158,SE,21N,-158W,2.275,Cr
166,164,SE,21N,-151W,5.756,Cr
172,170,SE,17N,-151W,1.627,Cr
173,171,SE,17N,-151W,2.396,Cr
177,175,SE,19N,-156W,1.235,Cr
178,176,SE,19N,-156W,0.654,Cr
179,177,SE,19N,-156W,1.298,Cr
180,178,SE,21N,-158W,9.504,Cr
181,179,SE,21N,-158W,0.600,Cr
182,180,SE,19N,-156W,1.051,Cr
183,181,SE,21N,-159W,8.277,Cr
184,182,SE,24N,-157W,0.853,Cr
185,183,SE,20N,-156W,5.015,Cr
186,184,SE,21N,-158W,0.960,Cr
187,185,SE,20N,-156W,1.782,Cr
188,186,SE,21N,-158W,2.716,Cr
201,199,SE,25N,-154W,0.891,Cr
202,200,SE,8N,-145W,0.000,Cr
206,204,SE,8N,-145W,-1.000,Cr
208,206,SE,12N,-145W,0.000,Cr
211,210,SE,16N,-148W,4.788,Cr
215,214,SE,24N,-153W,0.482,Cr
223,222,SE,25N,-148W,0.297,Cr
```

Surface structure of water skaters

Finding Your Researchers and Learning Their Language

In Other Words, Data Training

Tim Dennis Saw A Gap

- Grad students desperately needed training in R
    - Tim did some research and found that Software Carpentry had training materials
    - Tim also found a trainer on campus that was willing to help him teach a workshop

Within two days, there was a 50+ person waiting list

A Faculty member was so Impressed, he gave Tim funding to become an institutional member of Software Carpentry & integrated the course into the curriculum
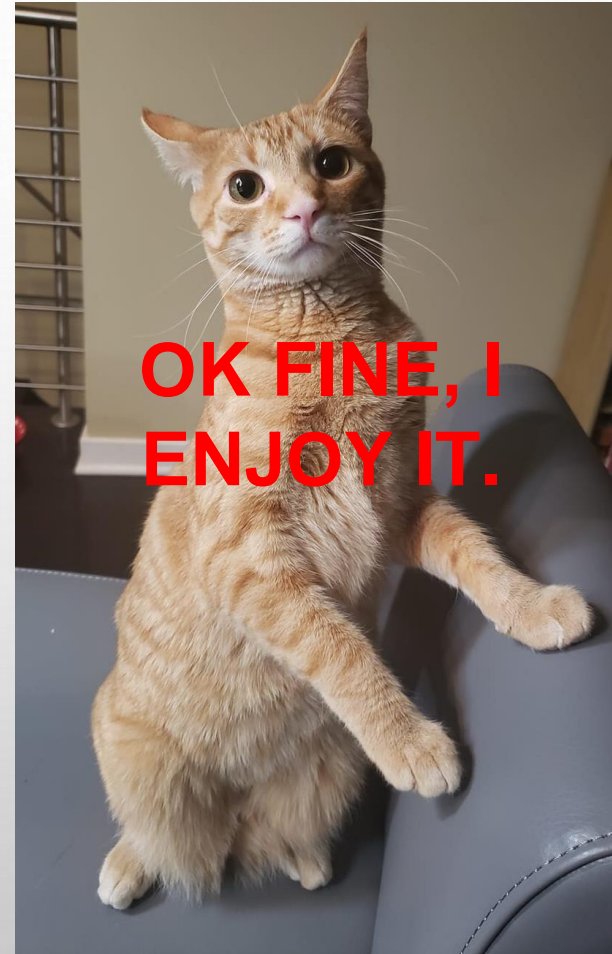
Not Satisfied With That, Tim Made Me Teach

And I found that:

- It strengthened my own data skills

- Kept me up to date with new developments

- Created productive relationships with researchers

I really loved it!

The Grand Conclusion!

- Every institution is different and has its strong points with regard to
  - Communication channels
  - Infrastructure
  - Policy

- Work with your strong points!

- Know what your mission and scope are
  - You can always expand those things, but if they start out too big, they are hard to reign back in

More Conclusions

- Find the need crevasses and fill or bridge them
  - You will be seen as collaborators and pain-easers
  - What seems like a small thing to you may substantially change a researcher's work experience
  - Start with the small things
- Be kind to yourself and recognize successes
  - They don't have to be major, obvious things

Thank you!

THANKS!

Questions?

Please put them in the chat box.

Slides and a recording will be sent to all registered delegates.