Dec 11th, 9:00 AM

Value of Curation

Lisa Johnston
University of Minnesota - Twin Cities, ljohnsto@umn.edu

The Value of Curation

Introduction to the workshop

Research data have value beyond their original purpose

How much data does the U have?

- The U has about 45PB of data we know of
- For comparison
  - The total hard drive space manufactured in 1995 was 20PB
  - The printed collection of the US Library of Congress is equal to about 10TB
  - The entire written works of mankind, from the beginning of recorded history, in all languages is 50PB*
- Most is medium to low security and either small/ archival or large/active

- Minnesota Supercomputing Inst.
  - About 10PB
    - 42% 150GB or less
    - 46% between 150GB and 5 TB
    - 10% between 5 and 20TB
    - 2% more than 20TB

*Source: https://siliconangle.com/blog/2013/11/13/how-big-is-big-data-really/

book and page numbers

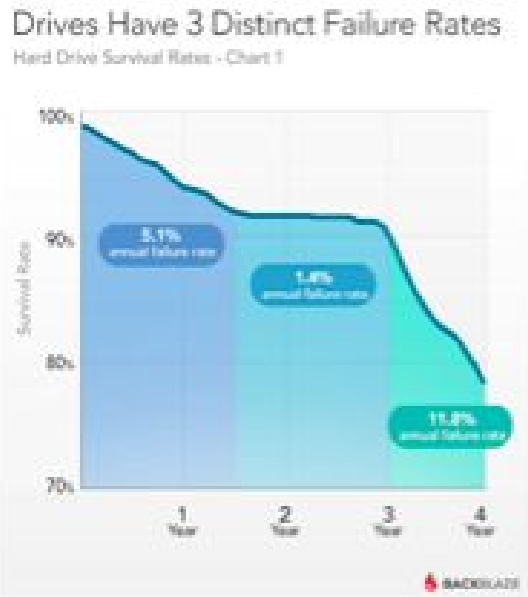
make indexing your work easier

sea_for_mfdfa.csv

A	B	C	D	E	F	G	H	I	J	K	L	M
0.296	-0.55	-0.243	-3.04	-2.885	0.071	1.847	-0.521	0.107	-3.528	-3.089	2.45	-
0.354	-0.487	-0.193	-2.974	-2.933	-0.029	1.951	-0.506	0.061	-3.643	-3.162	2.131	-1.
0.438	-0.449	-0.099	-2.979	-2.901	-0.09	2.051	-0.501	-0.005	-3.647	-3.34	1.713	-1.
0.519	-0.431	-0.018	-3.042	-2.831	-0.13	2.107	-0.452	-0.027	-3.616	-3.42	1.322	-1.
0.696	-0.37	0.023	-3.065	-2.832	-0.187	2.202	-0.415	-0.072	-3.648	-3.504	0.831	-1.
0.939	-0.332	0.083	-3.089	-2.815	-0.233	2.314	-0.342	-0.119	-3.603	-3.648	0.38	-1.
1.188	-0.295	0.171	-3.08	-2.753	-0.295	2.431	-0.197	-0.186	-3.598	-3.619	-0.07	-1.
1.503	-0.284	0.279	-3.129	-2.746	-0.363	2.52	-0.116	-0.298	-3.487	-3.5	-0.608	-1.
1.826	-0.288	0.36	-3.183	-2.743	-0.496	2.59	-0.012	-0.316	-3.318	-3.456	-0.989	-1.
2.153	-0.289	0.369	-3.162	-2.632	-0.615	2.653	0.197	-0.345	-3.249	-3.388	-1.33	-0.
2.59	-0.244	0.359	-3.205	-2.51	-0.761	2.761	0.412	-0.416	-3.204	-3.296	-1.58	-0.
2.97	-0.196	0.319	-3.218	-2.463	-0.944	2.933	0.643	-0.421	-3.143	-3.089	-1.746	-0.
3.269	-0.222	0.297	-3.148	-2.454	-1.045	3.051	0.904	-0.356	-2.983	-2.829	-1.813	-
3.512	-0.266	0.274	-3.157	-2.429	-1.147	3.119	1.116	-0.286	-2.783	-2.595	-1.927	-0.
3.684	-0.271	0.289	-3.214	-2.396	-1.255	3.052	1.222	-0.227	-2.627	-2.292	-2.081	-0.
3.824	-0.275	0.233	-3.289	-2.4	-1.262	2.996	1.39	-0.16	-2.475	-2.019	-2.286	-0.
3.889	-0.294	0.186	-3.295	-2.303	-1.306	2.961	1.545	-0.083	-2.293	-1.825	-2.461	-0.
3.896	-0.295	0.158	-3.289	-2.266	-1.383	2.93	1.645	-0.095	-2.195	-1.606	-2.573	-0.
3.838	-0.283	0.152	-3.286	-2.273	-1.352	2.876	1.615	-0.074	-2.086	-1.385	-2.672	-0.
3.712	-0.338	0.139	-3.328	-2.23	-1.302	2.778	1.637	-0.007	-1.971	-1.328	-2.759	0.
3.526	-0.363	0.125	-3.387	-2.198	-1.275	2.604	1.624	-0.019	-1.814	-1.35	-2.817	0.

Preventing Data Corruption

- Do backups
- Google Drive (cloud storage)
- Reliable hardware
  - Backblaze kept up to 25,000 hard drives constantly online for four years. Every time a drive fails, they note it down, then slot in a replacement.
  - 80% of drives last four years
  - https://www.extremetech.com/computing/170748-how-long-do-hard-drives-actually-live-for



Res nal
pur

- T
- D
- D
- N

GLOBAL DATA SHARING TRENDS

Data sharing practices vary widely across research fields and geographic areas. Just over half of researchers report making their data publicly available, though archiving results in repositories is not yet the norm.

WAYS DATA IS SHARED

- **67%** As supplementary material in a journal
- **37%** Personal, institutional or project webpage
- **26%** Institutional data repository *(i.e. university or institute-sponsored)*
- **19%** Discipline-specific data repository
- **6%** General-purpose data repository *(e.g. Dryad, figshare)*
- ✓ **5%** Other

Globally, researchers also report sharing their data in limited and non-permanent ways: 57% are sharing data at a conference while 42% of researchers share their data upon informal request (e.g. email, direct contact, etc.).

RESEARCHER MOTIVATIONS FOR SHARING

- 57% Data sharing is standard practice within my research community
- 55% To increase the impact and visibility of my research
- 50% Public benefit
- 42% Journal requirement
- 37% Transparency and re-use

DATA SHARING TRENDS BY COUNTRY

46% SHARING
54% NOT SHARING

UNITED STATES
Among researchers in the US sharing their data publicly, two out of three do so because it is standard practice in the communities and because they believe it benefits the public. Similar to their counterparts in the UK, the majority of US-based researchers also share data to increase the impact or visibility of their research.

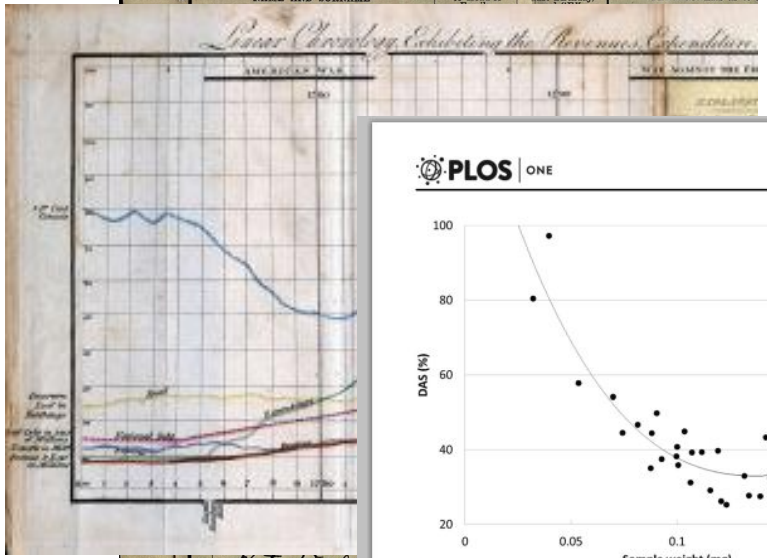
43% SHARING
57% NOT SHARING

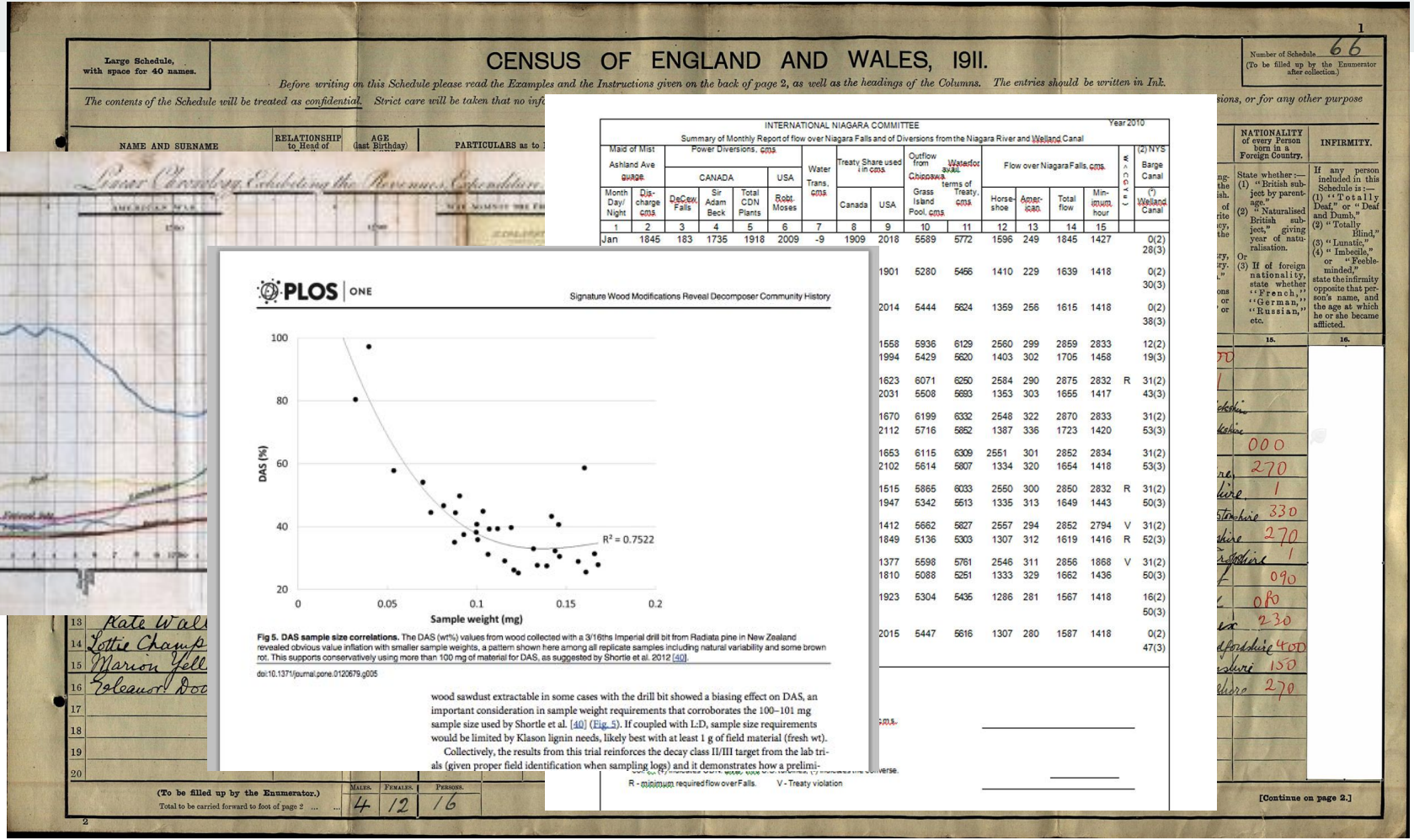
UNITED KINGDOM
While more than 40% of UK researchers are sharing data, only about 14% are using discipline-specific or other public repositories like Dryad and figshare. The two key drivers that motivate UK researchers to share their data are the prospect of gaining increased impact or visibility for their work and to satisfy funder requirements.

44% SHARING
56% NOT SHARING

JAPAN
Compared with their counterparts around the world, researchers in Japan cite concerns about being scooped as a reason for not sharing data more frequently. Nearly five out of ten Japanese researchers point to this as a reason for not sharing their data, roughly double the global average.

Source: Wiley's Researcher Data Insights 2014 Survey of 2,250 responses from researchers worldwide.





CENSUS OF ENGLAND AND WALES, 1911.

Large Schedule, with space for 40 names.

Number of Schedule 66 (To be filled up by the Enumerator after collection.)

Before writing on this Schedule please read the Examples and the Instructions given on the back of page 2, as well as the headings of the Columns. The entries should be written in Ink.

The contents of the Schedule will be treated as confidential. Strict care will be taken that no inf... ...sions, or for any other purpose

NAME AND SURNAME	RELATIONSHIP to Head of	AGE (last Birthday)	PARTICULARS as to	NATIONALITY of every Person born in a Foreign Country.	INFIRMITY.
				State whether:— (1) "British subject by parentage." (2) "Naturalised British subject," giving year of naturalisation. Or (3) If of foreign nationality, state whether "French," "German," "Russian," etc.	If any person included in this Schedule is:— (1) "Totally Deaf," or "Deaf and Dumb," (2) "Totally Blind," (3) "Lunatic," (4) "Imbecile," or "Feeble-minded," state the infirmity opposite that person's name, and the age at which he or she became afflicted.
				15.	16.
13 Kate Wals				000	
14 Lottie Champ				270	
15 Marion Yell				1	
16 Eleanor Doo				330	
17				270	
18				1	
19				090	
20				080	
				230	
				400	
				150	
				270	

(To be filled up by the Enumerator.) Total to be carried forward to foot of page 2 ... ...

Males.	Females.	Persons.
4	12	16

[Continue on page 2.]

INTERNATIONAL NIAGARA COMMITTEE — Year 2010

Summary of Monthly Report of flow over Niagara Falls and of Diversions from the Niagara River and Welland Canal

Maid of Mist / Ashland Avenue gauge		Power Diversions, cms				Water Trans, cms	Treaty Share used in cms		Outflow from Chippawa Grass Island Pool, cms	Waterfor avail. terms of Treaty, cms	Flow over Niagara Falls, cms				Remarks	(2) NYS Barge Canal (*) Welland Canal
		CANADA			USA											
Month Day/ Night	Discharge cms	DeCew Falls	Sir Adam Beck	Total CDN Plants	Robt. Moses		Canada	USA			Horseshoe	American	Total flow	Minimum hour		
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
Jan	1845	183	1735	1918	2009	-9	1909	2018	5589	5772	1596	249	1845	1427		0(2) 28(3)
								1901	5280	5456	1410	229	1639	1418		0(2) 30(3)
								2014	5444	5624	1359	256	1615	1418		0(2) 38(3)
								1558	5936	6129	2560	299	2859	2833		12(2)
								1994	5429	5620	1403	302	1705	1458		19(3)
								1623	6071	6250	2584	290	2875	2832	R	31(2)
								2031	5508	5693	1353	303	1655	1417		43(3)
								1670	6199	6332	2548	322	2870	2833		31(2)
								2112	5716	5852	1387	336	1723	1420		53(3)
								1653	6115	6309	2551	301	2852	2834		31(2)
								2102	5614	5807	1334	320	1654	1418		53(3)
								1515	5865	6033	2550	300	2850	2832	R	31(2)
								1947	5342	5513	1335	313	1649	1443		50(3)
								1412	5662	5827	2557	294	2852	2794	V	31(2)
								1849	5136	5303	1307	312	1619	1416	R	52(3)
								1377	5598	5761	2546	311	2856	1868	V	31(2)
								1810	5088	5251	1333	329	1662	1436		50(3)
								1923	5304	5436	1286	281	1567	1418		16(2) 50(3)
								2015	5447	5616	1307	280	1587	1418		0(2) 47(3)

R - minimum required flow over Falls. V - Treaty violation

PLOS | ONE

Signature Wood Modifications Reveal Decomposer Community History

Fig 5. DAS sample size correlations. The DAS (wt%) values from wood collected with a 3/16ths Imperial drill bit from Radiata pine in New Zealand revealed obvious value inflation with smaller sample weights, a pattern shown here among all replicate samples including natural variability and some brown rot. This supports conservatively using more than 100 mg of material for DAS, as suggested by Shortle et al. 2012 [40].

doi:10.1371/journal.pone.0120679.g005

wood sawdust extractable in some cases with the drill bit showed a biasing effect on DAS, an important consideration in sample weight requirements that corroborates the 100–101 mg sample size used by Shortle et al. [40] (Fig. 5). If coupled with L:D, sample size requirements would be limited by Klason lignin needs, likely best with at least 1 g of field material (fresh wt).

Collectively, the results from this trial reinforces the decay class II/III target from the lab trials (given proper field identification when sampling logs) and it demonstrates how a prelimi-

Some journals require data sharing

Example policy from PLOS

- *Make all data underlying the findings described in their manuscript fully available without restriction.*
- *When submitting a manuscript online, authors must provide a Data Availability Statement.*
- *Refusal to share data and related metadata and methods in accordance with this policy will be grounds for rejection.*
- *Methods acceptable to PLOS journals with respect to data sharing are:*
  - *Deposit data into appropriate repository (strongly recommended).*
  - *Include data in Supporting Information files.*
  - *Data made available to all interested researchers upon request.*
  - *Data available from third party. The reasons for restrictions on public data deposition must be specified.*

PLOS: http://www.plosone.org/static/policies#sharing

Incentives for sharing data strengthen the need for better curation

- Increased digital connectivity world-wide
- Funding mandates (e.g., NSF, Gates) for Data Management Plan (DMPs)
  - address how research will be “publicly accessible to search, retrieve, and analyze” (Holdren, 2013)
- Publisher data sharing policies (PLoS and Nature)
- “Reproducibility crisis” => standardized practices around data pipelines and replication studies
- Retraction Watch (stick) and Open Data movement (carrot) helps safeguard against scientific fraud or the dissemination of erroneous results

What is data curation?

Data curation is the active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education; curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time. (UIUC, 2007)

- Based in archival best practice (libraries know how to do this!)
- Data repositories provide a technological foundation
- But many curation activities are not easily automated ⇒ need curators (people)

Well-curated data are...

- Easier for fellow scholars and future collaborators to understand
- More likely to be trusted
- The research they represent are more likely to be reproducible
- More likely to be properly cited
- Represent potential cost-savings
- Findable, accessible, interoperable, and reusable, or FAIR (Wilkinson et. al, 2016)

Role of libraries in data curation

Libraries and academic-based data repositories are just one piece of the data repository landscape.

TABLE 4.3
Examples of kinds of data repositories found in the United States.

Kind of Repository	Examples
Federally Funded Data Centers	NASA Distributed Active Archives (DAAC), NOAA National Centers for Environmental Information (NCEI), National Snow and Ice Data Center (NSIDC), USGS Earth Resources Observation Systems (EROS) Data Center (EDC)
Federally Funded Research and Development Centers (FFRDC)	National Center for Atmospheric Research (NCAR), Jet Propulsion Lab (JPL), Oak Ridge National Laboratory (ORNL)
National Libraries	National Library of Medicine (NLM), National Agricultural Library (NAL), Library of Congress (LOC)
State and Local Agencies	State geological surveys, County planning offices
Thematic Repository	Long Term Ecological Research Network Information System (LTER NIS), Andrews Forest LTER (AND), National Snow and Ice Data Center (NSIDC), Maria Rogers Oral History Program
Domain Repository	Global Biodiversity Information Facility (GBIF), Inter-university Consortium for Political and Social Research (ICPSR), DataOne, Interdisciplinary Earth Data Alliance (IEDA)
Institutional Repository	Purdue University Research Repository (PURR), Data Repository for the University of Minnesota (DRUM)
Replication Repository	Dryad Digital Repository, Pangaea Data Library
Software Repository	GitHub, SourceForge
Commercial Archives	DigitalGlobe, Aerial photography companies, Resource exploration companies, Figshare
Private Archives	Huntington Library, Getty Research Institute

Baker, K. and Duerr, R. (2017). “Data and a diversity of repositories” in *Curating Research Data: A handbook of current practice* (L. R. Johnston, ed.). ACRL press.

Current state of libraries and data curation

- Surveyed 124 Association of Research Libraries (ARL) institutions in January 2017
- 80 institutions (65%) responded
- Goal: Understand the current data curation services offered and level of demand.

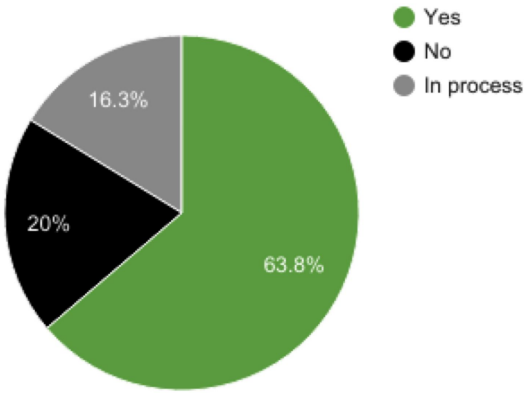
Cynthia Hudson-Vitale, Heidi Imker, Lisa R. Johnston, Jake R. Carlson, Wendy Kozlowski, Robert K. Olendorf, and Claire Stewart. (2017). SPEC Kit #354: Data Curation. Association of Research Libraries (ARL). May 2017. http://publications.arl.org/Data-Curation-SPEC-Kit-354/~~FreeAttachments/Data-Curation-SPEC-Kit-354.pdf

Result: Nearly two thirds (51/80) provided data curation services, 13 planning services, 16 not

Of those that provided data curation services:

- **Recent/new service**: More than half began in 2010 or later.
- **Repository-focused**: Nearly all also provided repository services
- **Based in IR**: More than half had an institutional repository that accepted data and a few had a stand-alone data repository.
- **Platforms** ranged from: DSpace (22), Fedora/Hydra (10), Islandora (7), Custom solution (7), Dataverse (local installation) (7), Digital Commons/BePress (5), Dataverse (hosted) (4), Other platform (10) such as HUBzero, Open Science Framework, Rosetta, and SobekCM.

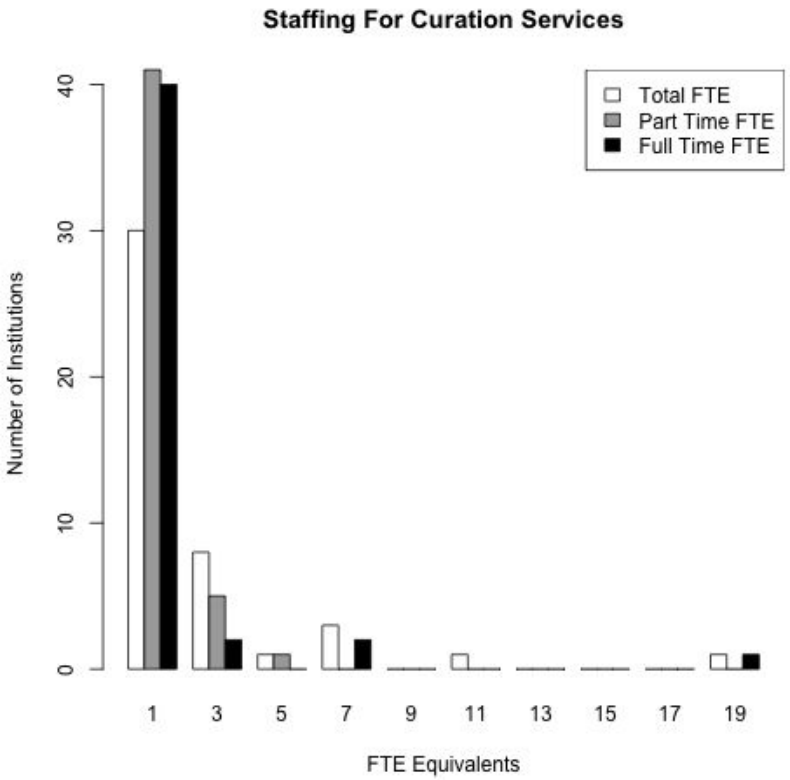
Does your institution currently provide research data curation services?



Cynthia Hudson-Vitale, Heidi Imker, Lisa R. Johnston, Jake R. Carlson, Wendy Kozlowski, Robert K. Olendorf, and Claire Stewart. (2017). SPEC Kit #354: Data Curation. Association of Research Libraries (ARL). May 2017. http://publications.arl.org/Data-Curation-SPEC-Kit-354/~~FreeAttachments/Data-Curation-SPEC-Kit-354.pdf

Staffing for Data Curation

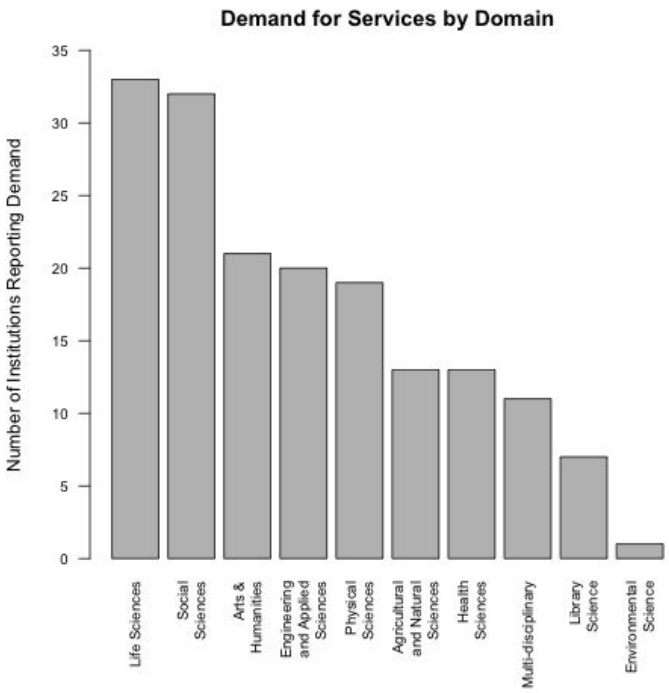
- Total Full-time equivalent (FTE) averaged to one person per institution dedicated to data curation services.
- Most libraries had 1 or more individuals providing data curation services at 5%-50% of their time while also carrying out other duties (part time FTE).



Cynthia Hudson-Vitale, Heidi Imker, Lisa R. Johnston, Jake R. Carlson, Wendy Kozlowski, Robert K. Olendorf, and Claire Stewart. (2017). SPEC Kit #354: Data Curation. Association of Research Libraries (ARL). May 2017. http://publications.arl.org/Data-Curation-SPEC-Kit-354/~~FreeAttachments/Data-Curation-SPEC-Kit-354.pdf

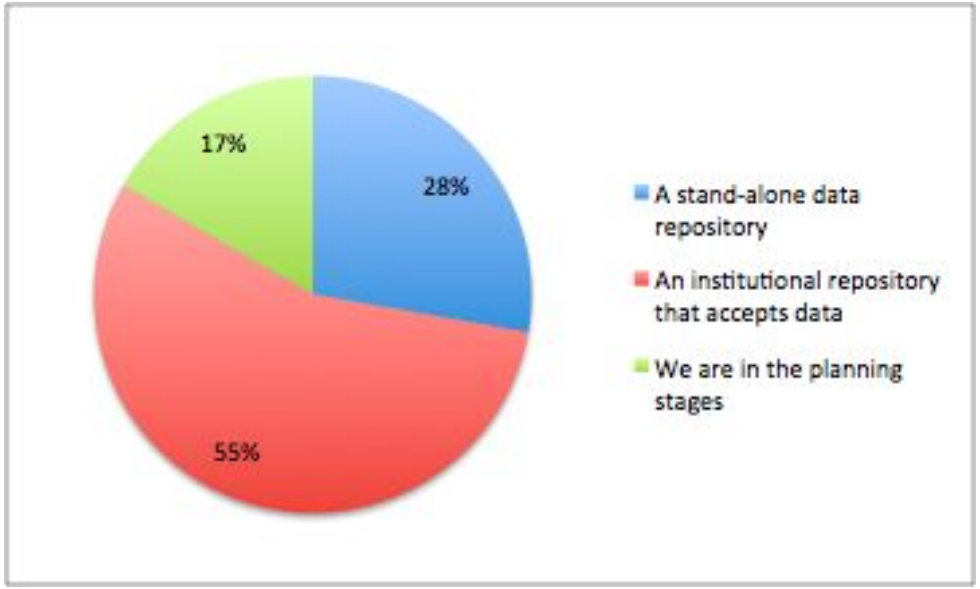
Demand for Data Curation

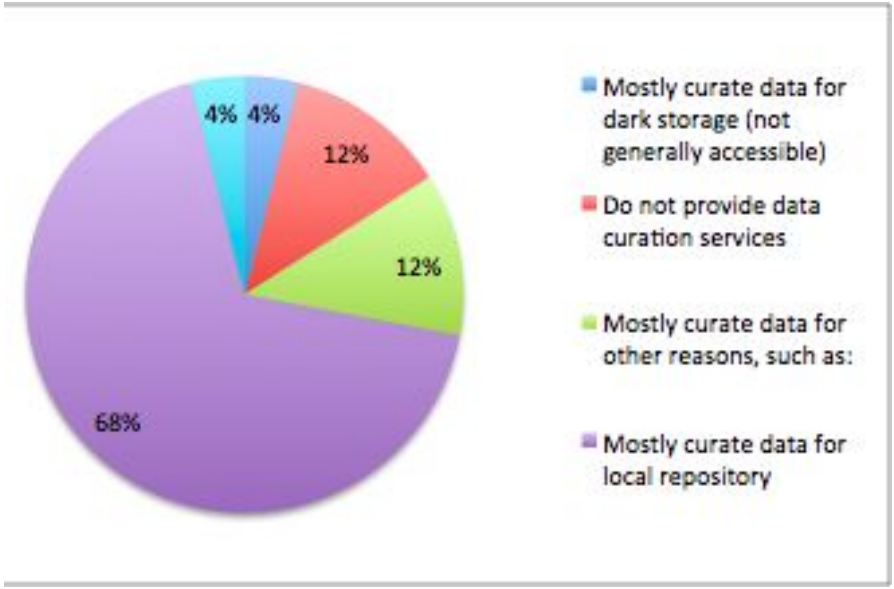
- **Domain:** Sciences dominated overall with Bio most freq, then Social Sci.
- **Frequency**: Most received only 1 dataset per month with 3 ingesting > 10/month.
- **Collection**: Of the libraries that had ingested data, ~half (26/46) had fewer than 50 data sets in their entire collection. Only 7 libraries had over 200 data sets in their data collection.



Cynthia Hudson-Vitale, Heidi Imker, Lisa R. Johnston, Jake R. Carlson, Wendy Kozlowski, Robert K. Olendorf, and Claire Stewart. (2017). SPEC Kit #354: Data Curation. Association of Research Libraries (ARL). May 2017. http://publications.arl.org/Data-Curation-SPEC-Kit-354/~~FreeAttachments/Data-Curation-SPEC-Kit-354.pdf

State of data curation at your institutions





Elevator Speech

An elevator speech should a 30-60 sec statement that includes one or more of the following:

(1) Identifies a goal: “We should implement data curation services in the next year...”

(2) Explains what you do: “Let me tell you about our data curation services...”

(3) Communicates your Unique Selling Proposition: “Libraries are great at this because…”

(4) Engages with a question “How hard is it for you to share your data…?”

Exercise: Choose an audience (e.g, someone from your library administration, faculty members, university admin, etc.) and draft a short statement **explaining the value of data curation and why it merits increased consideration or investment.**

Elevator Speech

Share your elevator speech with the attendees by uploading it to the shared notes folder

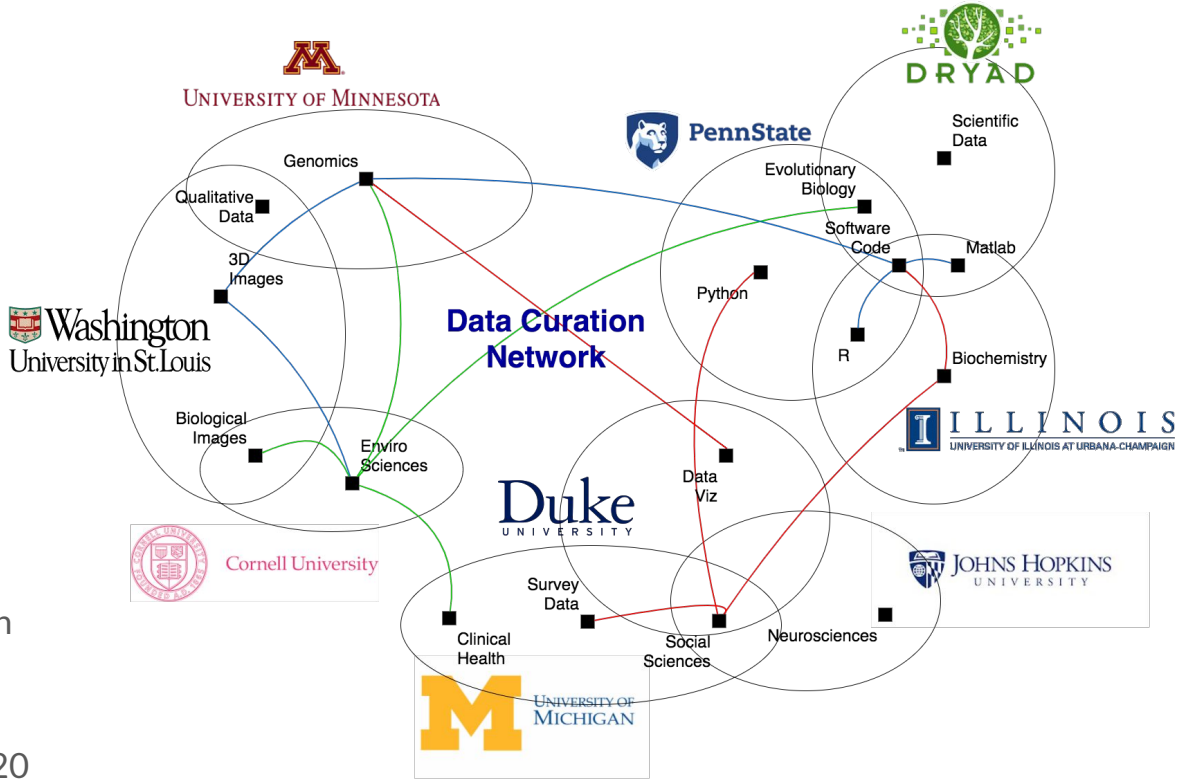
http://bit.ly/2klXVyq

Data Curation Activities

Research performed by the Data Curation Network

Data Curation Network

- Collaborative staffing model for curating research data
- Launched in 2016 with six institutions.
- One year planning 16-17 phase funded by Sloan Foundation
- Implementation phase (aiming for May 2018) will pilot the model with nine partner institutions
- Goal is to expand to all users in 2020



What Curation Activities are Important?

DCN Researcher Study 2016 ⇒ identify where to invest focus of DCN

Method: Held focus groups (Oct-Nov 2016) at the 6 DCN partner institutions, asked researchers:

1. How important are data curation activities for your data?
2. What data curation activities are currently being done by you or a 3rd party?
3. If the data curation activity is being performed, how satisfied are you with the results?

Full list of activities: http://bit.ly/DCNcurationActivities

Data Curation Activities

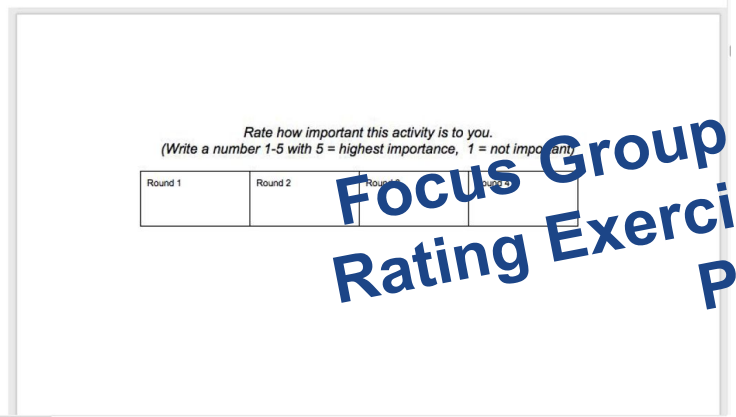
Code review
Contextualize
Documentation
Embargo
File Format Transformations
Persistent Identifier
Quality Assurance
Use Analytics

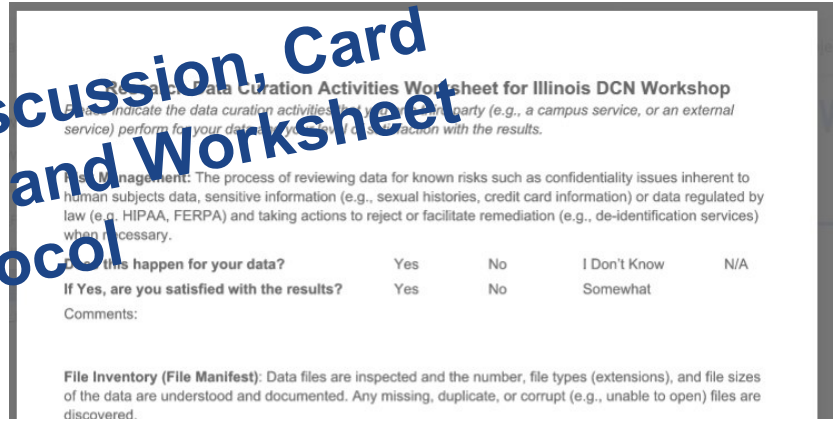
File re
File Inventory or Manifest
File validation
Metadata
Metadata Brokerage
Rights Management
Risk Management
……more

We identified and defined 47 Data Curation Activities

Sources: CASRAI Dictionary, Research Data Alliance (RDA) Terms Definition Tool, Digital Curation Center (DCC) Glossary, [book] *Curating Research Data: Handbook of Current Practice,"* ICPSR Glossary of Social Science Terms, Research Data Canada Glossary, Digital Preservation Coalition Glossary, and Society of American Archivists Terms Glossary.

Mixed Methods Approach





Focus Group Discussion, Card Rating Exercise, and Worksheet Protocol

[Preprint] Johnston, Lisa R., Jake R. Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert K. Olendorf, and Claire Stewart. 2017. "How Important is Data Curation? Gaps and Opportunities for Academic Libraries" Submitted for publication.

Exercise: Data Curation Activity Importance

1. Rate the data curation activity on each card.
2. Then trade with someone else in the room. (no repeats)
3. Repeat for 4 rounds per card

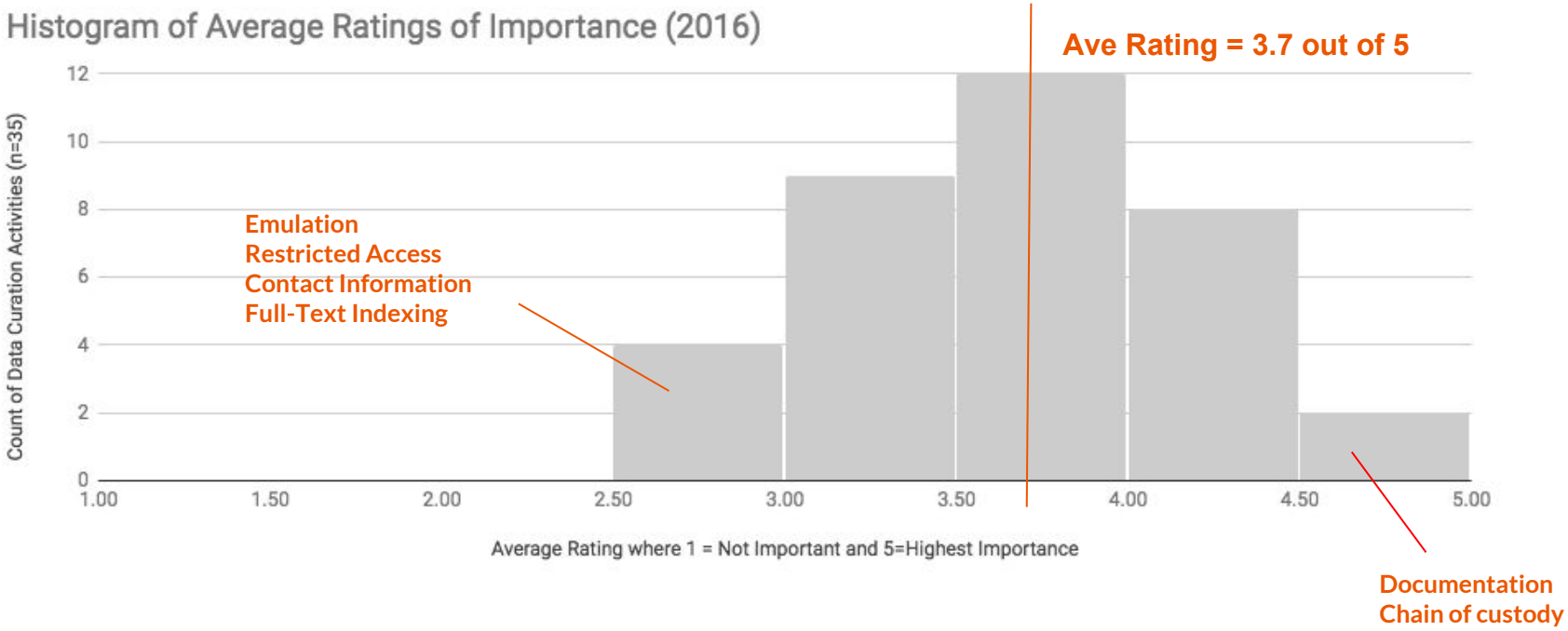
If you are the last to complete the card….

4. Total the ratings and calculate the average rating (divide total by 4)
5. Write the average rating on the FRONT of the card
6. Tape card to wall in order of ratings

Discussion: Data Curation Activities

1. Discusses the results of the exercise with others at your table.
2. What themes or surprises emerged?
3. How do our (library/archival) ratings of importance differ from researchers?

DCN Researcher Study 2016 (n=91)



DCN Researcher Study 2016 (n=91)

Most Important Activities* (4 out of 5)

- **(Create) Documentation (4.6)**
- **Secure Storage (4.4)**
- **Quality Assurance (4.3)**
- **Persistent Identifier (4.3)**
- **Software Registry (4.1)**
- **Data Visualization (4.0)**
- **File Audit (4.0)**
- **(Create) Metadata (4.0)**
- **Versioning (3.9)**
- **Contextualization (3.9)**
- **Code Review (3.9)**
- **File Format Transformations (3.9)**

* Rated by more than one DCN focus group from our 2016 Study

DCN Researcher Results 2016 (n=91)

Most Important Activities* (4 out of 5)

- (Create) Documentation (4.6)
- Secure Storage (4.4)
- Quality Assurance (4.3)
- Persistent Identifier (4.3)
- Software Registry (4.1)
- Data Visualization (4.0)
- File Audit (4.0)
- (Create) Metadata (4.0)
- Versioning (3.9)
- Contextualization (3.9)
- Code Review (3.9)
- File Format Transformations (3.9)

Not Happening for Majority of Researchers

- **Persistent Identifier** (37% happens)
- **Software Registry** (41% happens)
- **File Audit** (16% happens)
- **Contextualization** (38% happens)
- **Code Review** (38% happens)

* Rated by more than one DCN focus group from our 2016 Study

DCN Researcher Results 2016 (n=91)

Most Important Activities* (4 out of 5)

- (Create) Documentation (4.6)
- Secure Storage (4.4)
- Quality Assurance (4.3)
- Persistent Identifier (4.3)
- Software Registry (4.1)
- Data Visualization (4.0)
- File Audit (4.0)
- (Create) Metadata (4.0)
- Versioning (3.9)
- Contextualization (3.9)
- Code Review (3.9)
- File Format Transformations (3.9)

* Rated by more than one DCN focus group from our 2016 Study

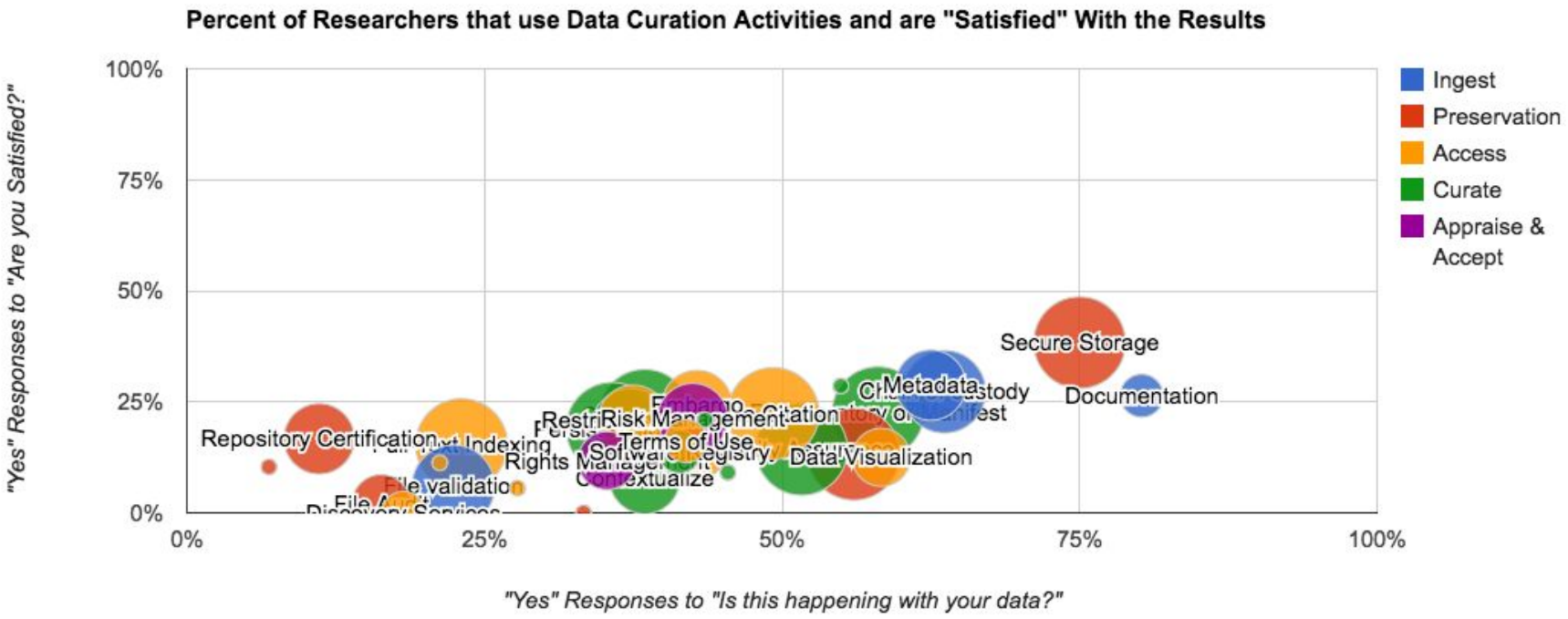
Not Happening for Majority of Researchers

- **Persistent Identifier** (37% happens)
- **Software Registry** (41% happens)
- **File Audit** (16% happens)
- **Contextualization** (38% happens)
- **Code Review** (38% happens)

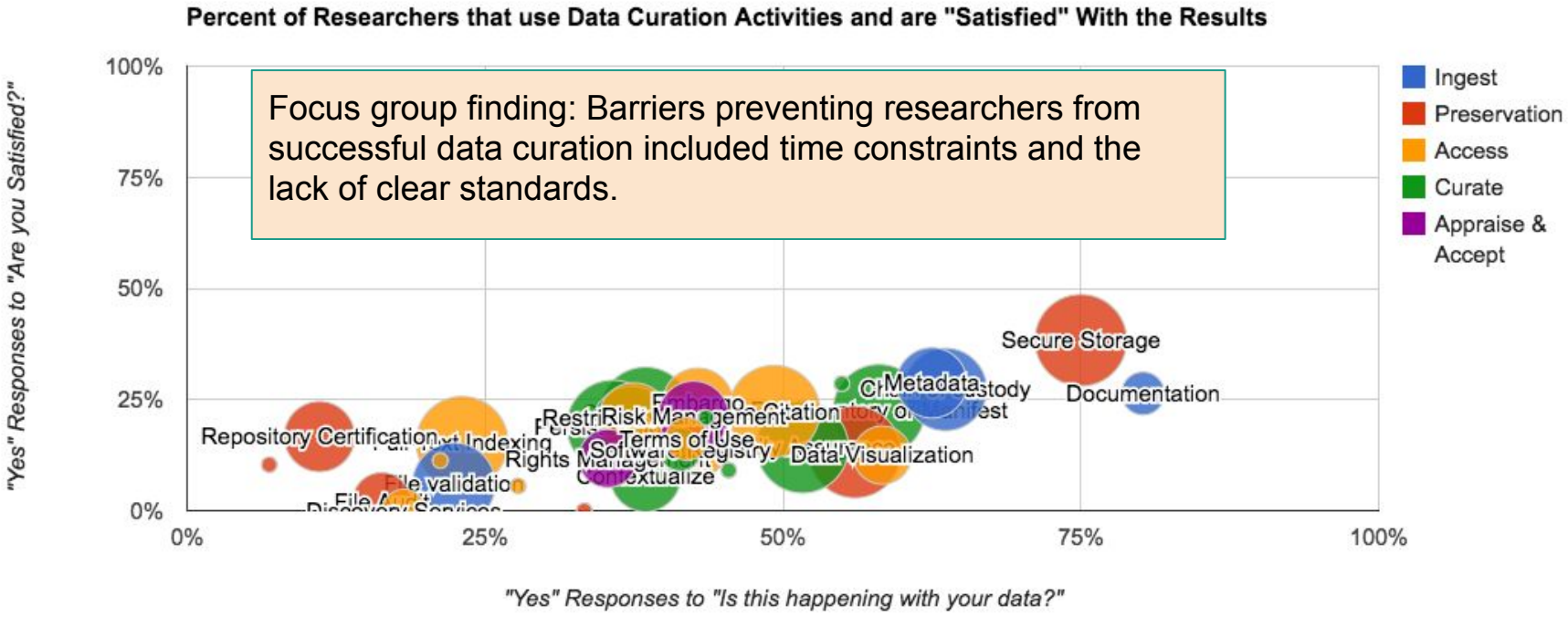
Happening, but not satisfactorily

- **Documentation** (26% satisfied),
- **Secure storage** (38% satisfied),
- **Quality Assurance** (14% satisfied),
- **Data Visualization** (12.5% satisfied),
- **Metadata** (29% satisfied)
- **Versioning** (13% Satisfied)
- **File Format Transformations** (29% satisfied)

Result: No Activity was Satisfying the Majority

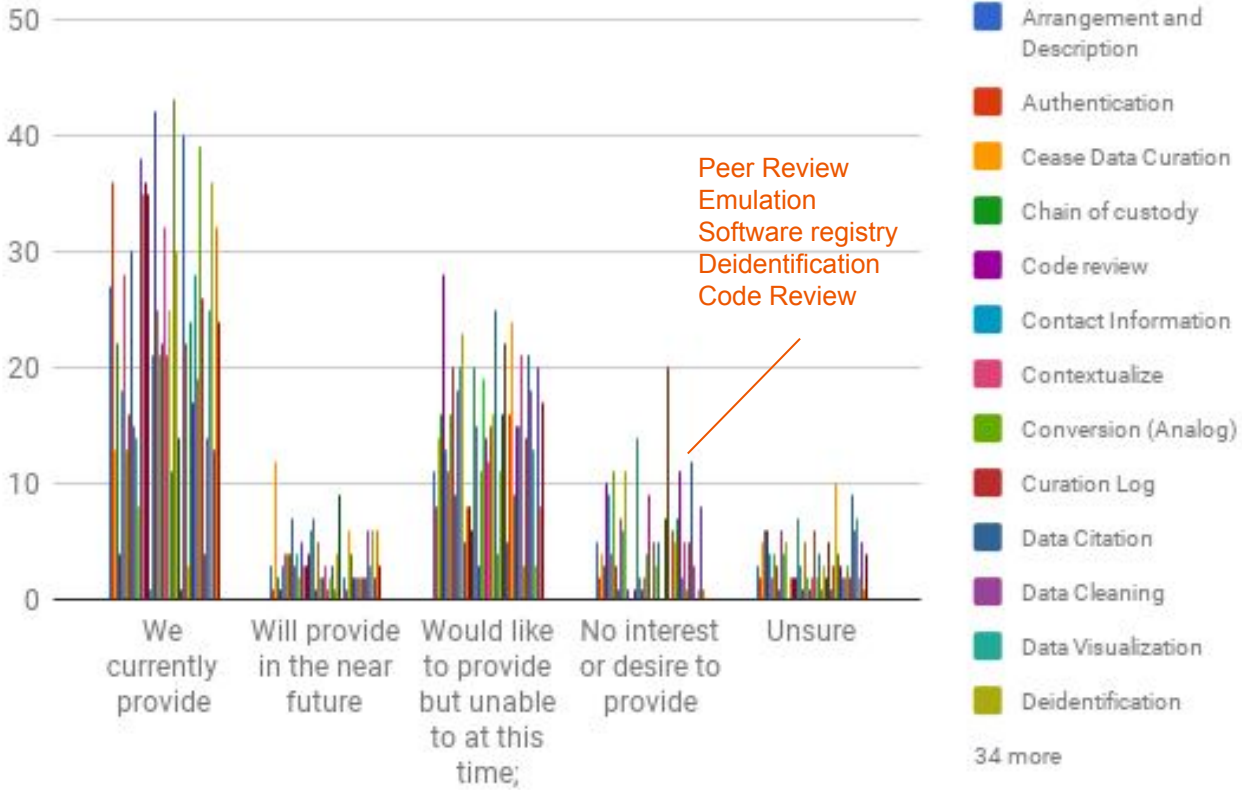


Result: No Activity was Satisfying the Majority



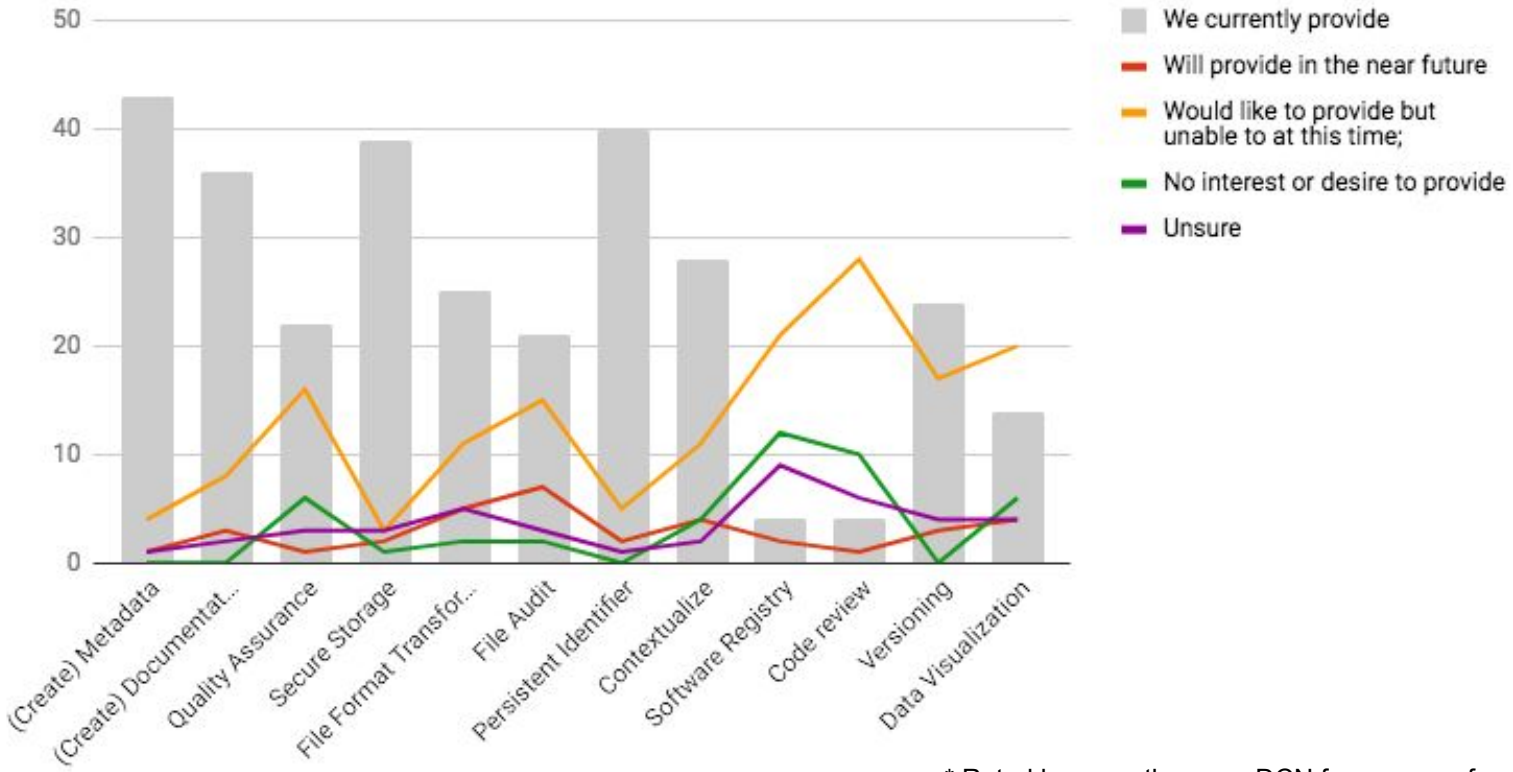
Focus group finding: Barriers preventing researchers from successful data curation included time constraints and the lack of clear standards.

SPEC Kit #354: We asked ARL Institutions to self-assess their support for 47 different data curation activities ranging from ingest activities to preservation actions.



Cynthia Hudson-Vitale, Heidi Imker, Lisa R. Johnston, Jake R. Carlson, Wendy Kozlowski, Robert K. Olendorf, and Claire Stewart. (2017). SPEC Kit #354: Data Curation. Association of Research Libraries (ARL). May 2017. http://publications.arl.org/Data-Curation-SPEC-Kit-354/~~FreeAttachments/Data-Curation-SPEC-Kit-354.pdf

How are the most important Data Curation Activities* supported at n=49 ARL institutions?



* Rated by more than one DCN focus group from our 2016 Study

Key Takeaways

- We need a shared language when discussing data curation activities
- Researchers are actively engaged data curation activities for their data
- Many activities may not be happening in a satisfactory way
- Libraries will most benefit from emphasizing, investing in, and/or heavily promoting the services that researchers value (rather than what we value)
- Gaps in satisfaction for highly-valued data curation activities provide opportunity for partnership.

Break

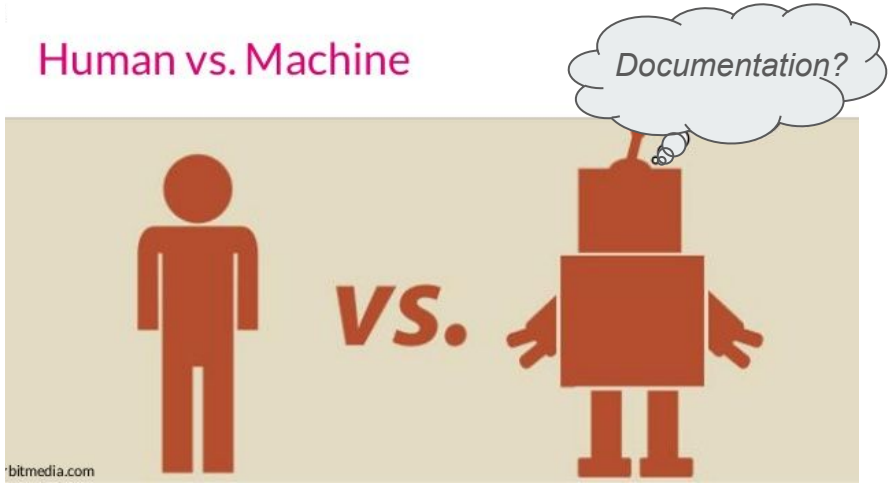
10:30-10:40

CURATE Model

CURATE Steps

Perform and document all actions taken in to....

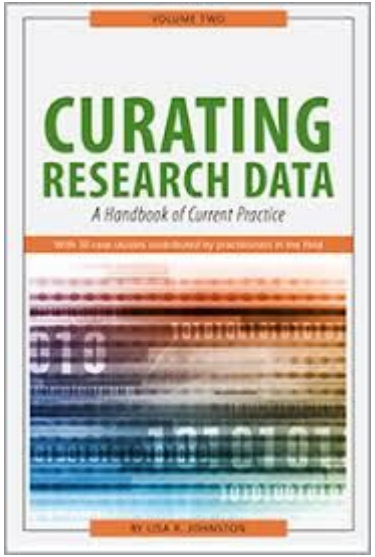
- **C** **Check** files and read documentation.
- **U** **Understand** the data (or try to), if not...
- **R** **Request** missing information or changes.
- **A** **Augment** metadata for findability.
- **T** **Transform** file formats for reuse.
- **E** **Evaluate** for FAIRness.



But first...things to consider before you CURATE

Establishing a data curation service involves several preliminary steps

1.



Johnston, Lisa R. "Curating Research Data Volume Two: A Handbook of Current Practice." (2017).Association of College and Research Libraries Press. Chicago, IL.

USGS
science for a changing world

Trusted Digital Repositories: What are they and how you become one

Contacts:
John Faundeen
Clara Brown
Keith Kirk

The **Trusted Digital Repository Working Group** is part of the USGS Fundamental Science Practices Advisory Committee (FSPAC) Data Preservation Subcommittee

How to apply for TDR Certification:

1. Obtain TDR application from TDR WG,
2. Organizational unit completes application,
3. Submits to TDR WG,
4. Submission Review,
   Keith Kirk, Clara Brown & John Faundeen (Representing FSPAC Data Preservation Subcommittee and TDR WG)
5. TDR status granted or submission returned for process modification,
6. Submission Status Spreadsheet
   - Maintained by TDR WG
   - Trigger for Re-Certification in Three Years

*"A **trusted digital repository** is one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future."*
(source: https://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf)

To obtain certification as a Trusted Digital Repository (TDR) the repository must meet the following criteria:

1. The repository has an explicit mission to provide access to and preserve data in its domain.
2. The repository maintains all applicable licenses covering data access and use and monitors compliance.
3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.
4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.
5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.
6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either in-house, or external, including scientific guidance, if relevant).
7. The repository guarantees the integrity and authenticity of the data.
8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.
9. The repository applies documented processes and procedures in managing archival storage of the data.
10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.
11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.
12. Archiving takes place according to defined workflows from ingest to dissemination.
13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.
14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.
15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.
16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

WG Members
Lance Everette
Ben Wheeler
Clara Brown Co-chair
David Boldt
John Faundeen Co-chair
Keith Richmond
Kelly Haberstroh
Natalie Latysh
Rex Sanders
Sofia Dabrowski
Tara Bell

Criteria Sources Reviewed:

- ✓ U.S. Federal RIM Maturity Model,
- ✓ Digital Curation Centre Checklist for Evaluating Data Repositories,
- ✓ NOAA Unified Framework,
- ✓ Data Seal of Approval,
- ✓ ISO 16363-2012 Module 8, Becoming a Trusted Digital Repository,
- ✓ LoC National Digital Stewardship Alliance,
- ✓ Data Seal of Approval / World Data System

USGS
Trusted Digital Repository

But
CU

Establi

1.

2.

3.

4.

5.

6.

7.

a.

Volume 3, Issue 2 (2015)

Do You Have an Institutional Data Policy? A Review of the Current Landscape of Library Data Services and Institutional Data Policies

Kristin Briney, Abigail Goben, Lisa Zilinski

Briney, K., Goben, A., & Zilinski, L. (2015). Do You Have an Institutional Data Policy? A Review of the Current Landscape of Library Data Services and Institutional Data Policies. *Journal of Librarianship and Scholarly Communication*, 3(2), eP1232. http://dx.doi.org/10.7710/2162-3309.1232

External Data or Supplements:

Briney, Kristin; Goben, Abigail; Zilinski, Lisa, 2015, "Data from: Do You Have an Institutional Data Policy? A Review of the Current Landscape of Library Data Services and Institutional Data Policies", http://dx.doi.org/10.7910/DVN/GAZPAJ, Harvard Dataverse

DCP

Home | Overview | Publish a Profile | DCP Directory | Archive | Contact

Download the Toolkit

Data Curation Profiles

Thank you for your interest and spreading the use of the ***Data Curation Profiles Toolkit***. In October 2014 we received an Institute of Museum and Library Services Planning Grant to continue working on on the DCP. The primary goal of this grant is to create a roadmap to scope the outcomes and work needed for a redesign. We believe this can contribute to curating more of research outputs—moving from active management of data and digital objects to dissemination and preservation of them. A "bridging the gap" report will be published later that pulls together the thinking of experts on issues and challenges for that. Stay tuned...

OECD publishing

BUSINESS MODELS FOR SUSTAINABLE RESEARCH DATA REPOSITORIES

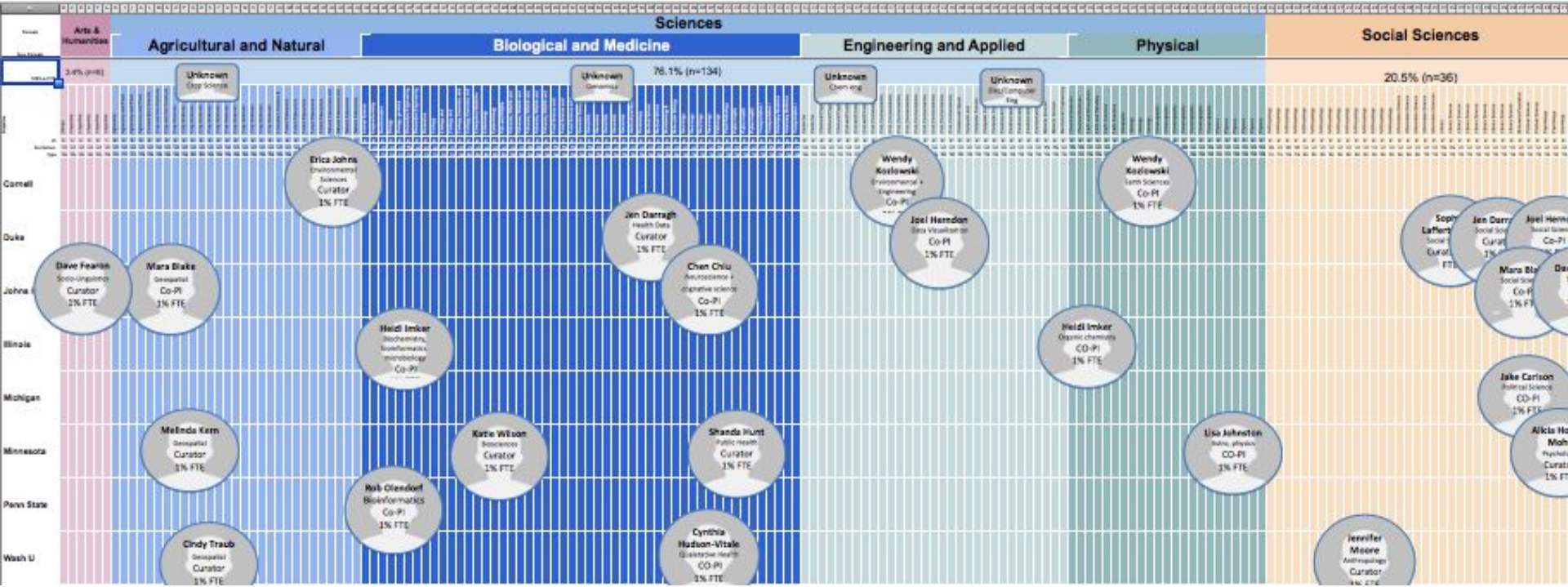
OECD SCIENCE, TECHNOLOGY AND INNOVATION
POLICY PAPERS

December 2017 **No. 47**

Data Archiving Infrastructure

Primary platform choice

Inst. Repository w/ Data (top 5)	Data-specific Repository
Dspace	Dataverse
Fedora	Chronopolis
BePress Digital Commons	HubZero (customized)
Hydra	DataConservancy
Drupal	Custom repository



7. Software: What specific tools and expertise will be utilized?

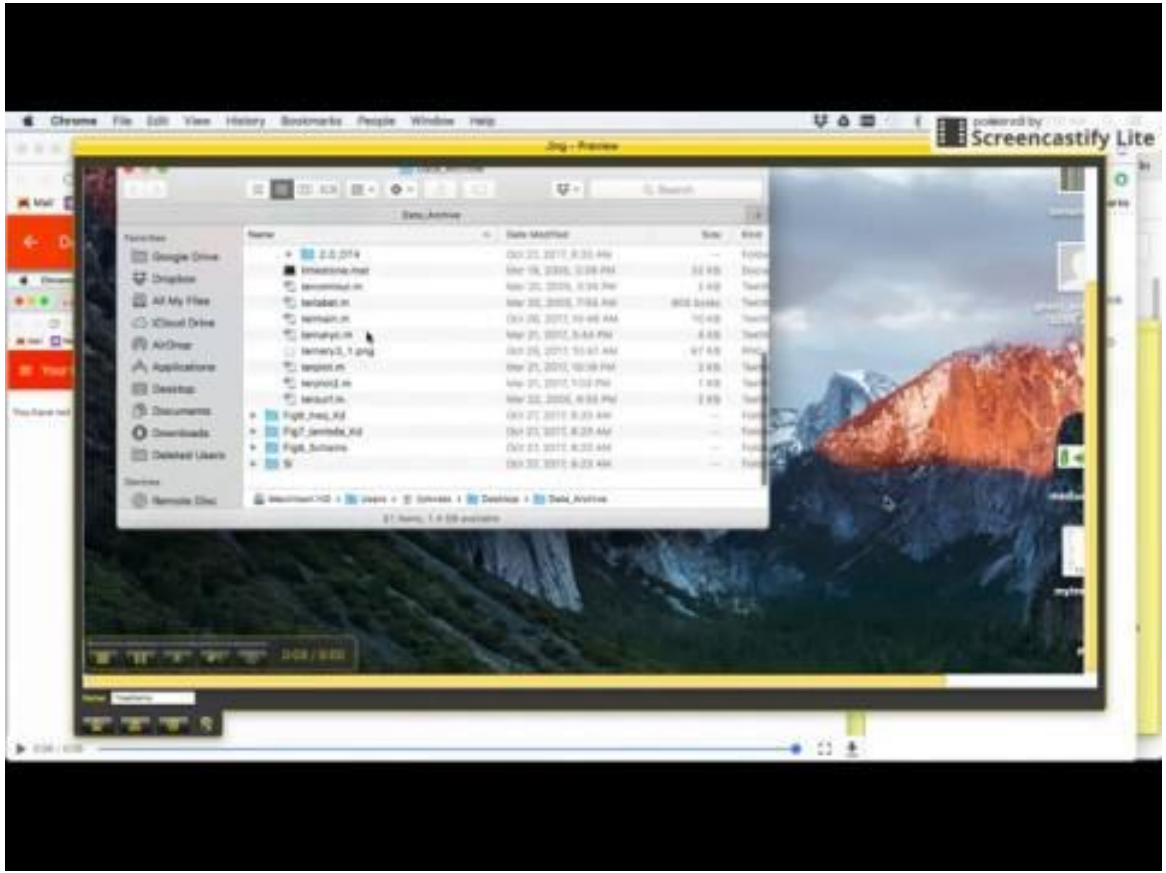
But first...things to consider before you CURATE

Establishing a data curation service involves several preliminary steps

1. Mission: What is the institutional commitment to providing data curation services at a level appropriate for your goals (e.g., is this a pilot?).
2. Policies: What data will be curated (e.g., criteria for acceptance and rejection).
3. Audience: Who are your stakeholders, what do they need?
4. Costs: How much will it cost (money, staff, time) to run your service?
5. Technology: What repository infrastructure will you use to securely ingest and store the data.
6. Staff: Who is involved and what is their expertise? Where are your gaps?
7. Software: What specific tools and expertise will be utilized?

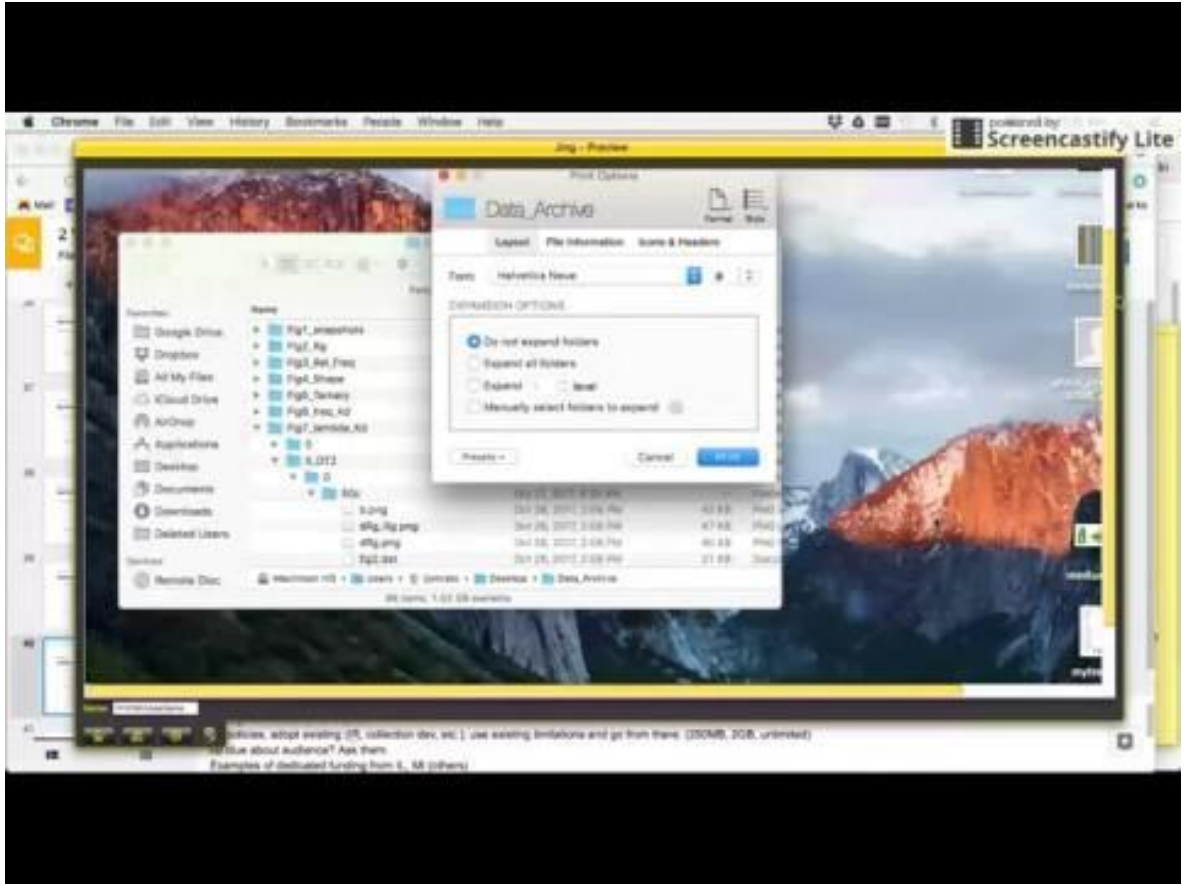
Software Roundup

1. Appraisal Tools
   a. **Tree**



Software Roundup

1. Appraisal Tools
   a. Tree and **Print Window (demo)**

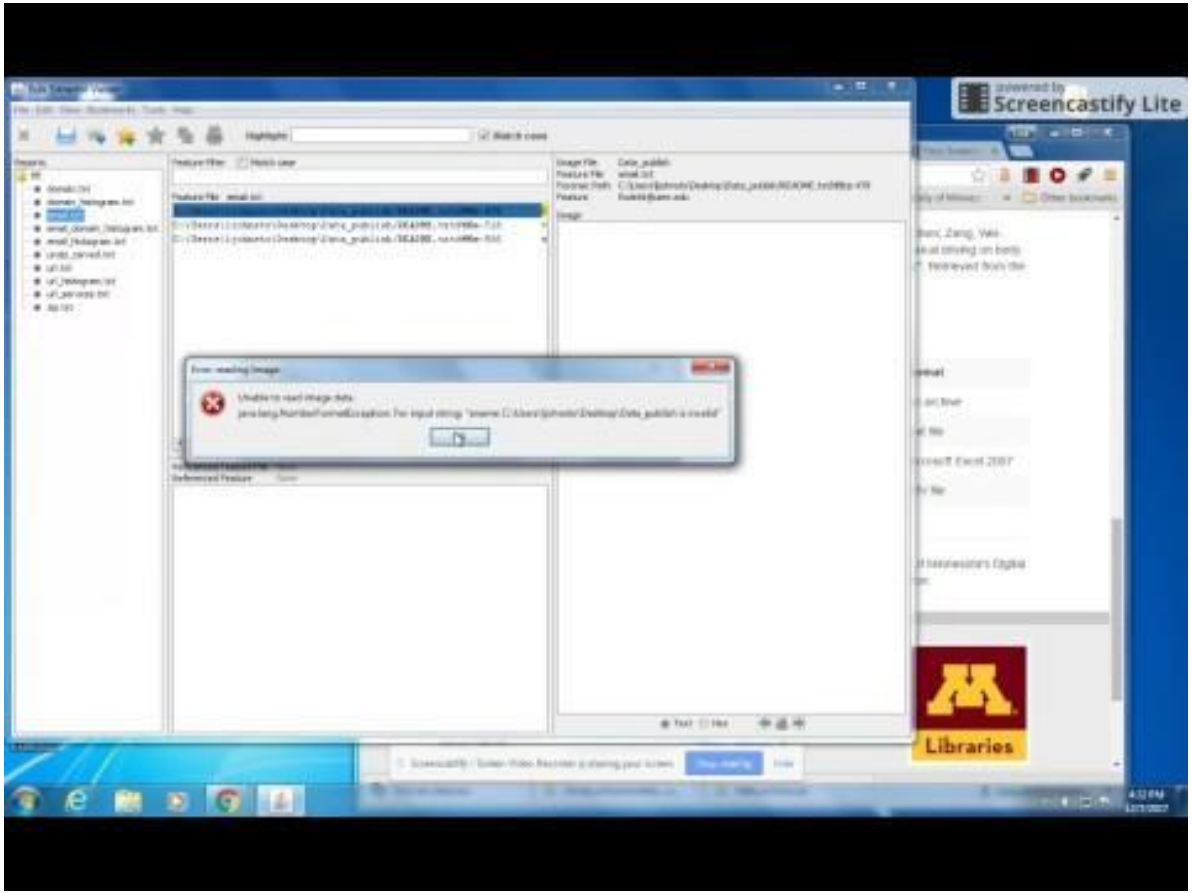


Software Roundup

1. Appraisal Tools
   a. Tree and Print Window (demo)
   b. **Identify Finder (demo)**

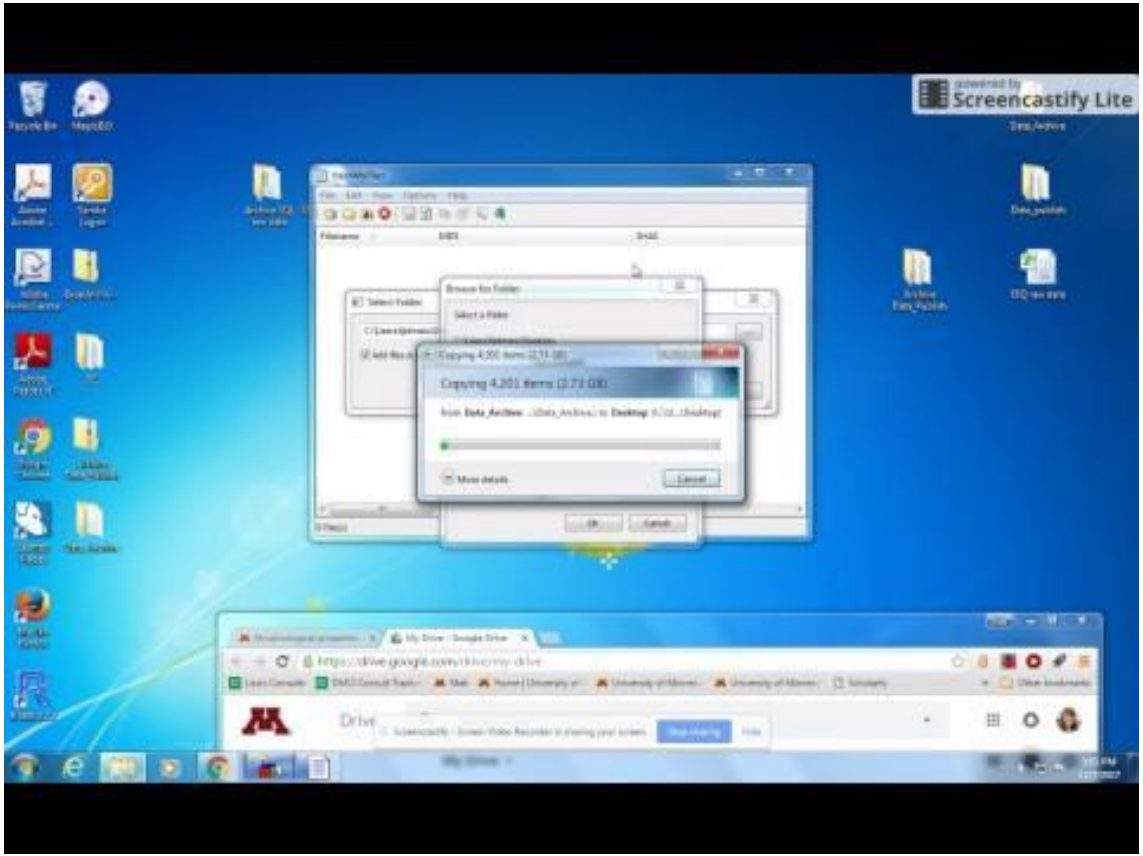
Software Roundup

1. Appraisal Tools
   a. Tree and Print Window (demo)
   b. Identify Finder (demo)
   c. **BulkExtractor**



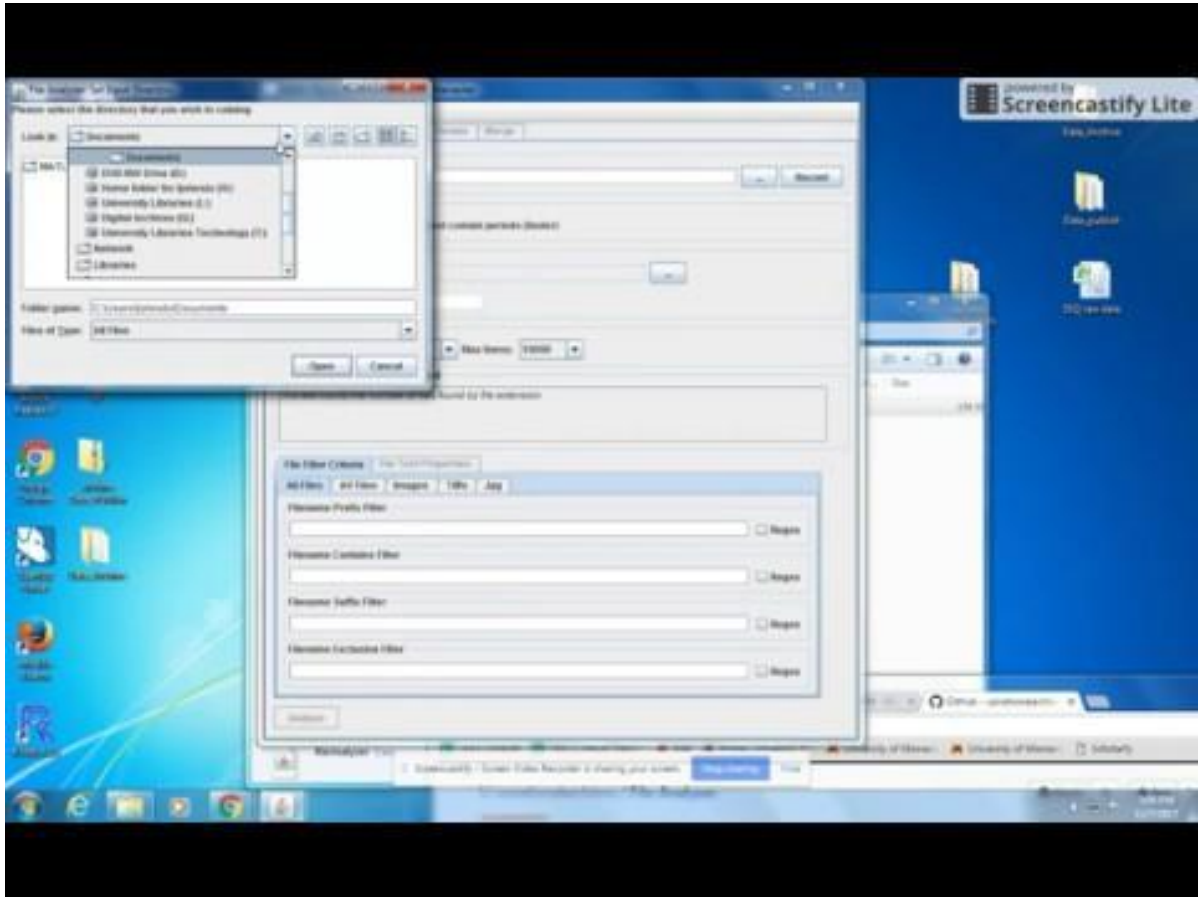
Software Roundup

1. Appraisal Tools
   - **a.** Tree and Print Window (demo)
   - b. Identify Finder (demo)
   - c. BulkExtractor
   - **d. HashMyFiles**



Software Roundup

1. Appraisal Tools
   - **a.** Tree and Print Window (demo)
   - b. Identify Finder (demo)
   - c. BulkExtractor
   - d. HashMyFiles
   - **e. NARA FIle Analyzer**



Software Roundup

1. Appraisal Tools
   a. Tree and Print Window (demo)
   b. Identify Finder (demo)
   c. BulkExtractor
   d. HashMyFiles
   e. NARA FIle Analyzer
2. Speciality Tools (Processing and Review)
   a. ArcGIS, QGIS, GeoNetwork
   b. RStudio
   c. Matlab
   d. AliView (genomics data viewer)
   e. Omero (microscopic images)
   f. ChemDraw
   g. MZmine - mass spec data
   h. Jena - 3d molecular structure/Crystallographic files

Discussion

What software/tools do you find useful for data curation?

1. Ditto (mac zip)
2. Fiji ImageJ
3. SosiriX (medical images)
4. NCBI image viewers
5. MeshLab
6. BulkRename Utility
7. Python data breakers (scripts)
8. EML metadata Morpho

Demo Datasets

Let’s meet our six case studies (5 mins each)

http://bit.ly/2kIXVyq