The Data Census: Assessing Data Services at MSU

Sara Mannheimer
Data Management Librarian, Montana State University
sara.mannheimer@montana.edu
@saramannheimer

CLIR/DLF E-Research Network
October 14, 2015

Survey Development

Collaboration between Library and Campus IT
(with input from Office of Planning and Analysis)

Informed by other institutional surveys and assessment frameworks
(see resources slide)

Survey Development

Size and types of research data

Data storage and backup behaviors

Data sharing behaviors

Survey Goal

Gauge interest in and need for institutional data curation, storage, and publication services

Distributing the Census

Distributed in January 2015

Email to academic faculty and staff was cosigned by Vice President for Research, Dean of the Library, and CIO

Vice President for Research, Dean of the Library, and CIO helped with publicity and encouraged responses

Qualtrics survey tool

Assumptions

Assumptions

Most researchers store their data digitally, on a personal computer or external hard drive

Researchers create a wide variety of data formats

Researchers have relatively small amounts of data—less than 100 Gb

Assumptions

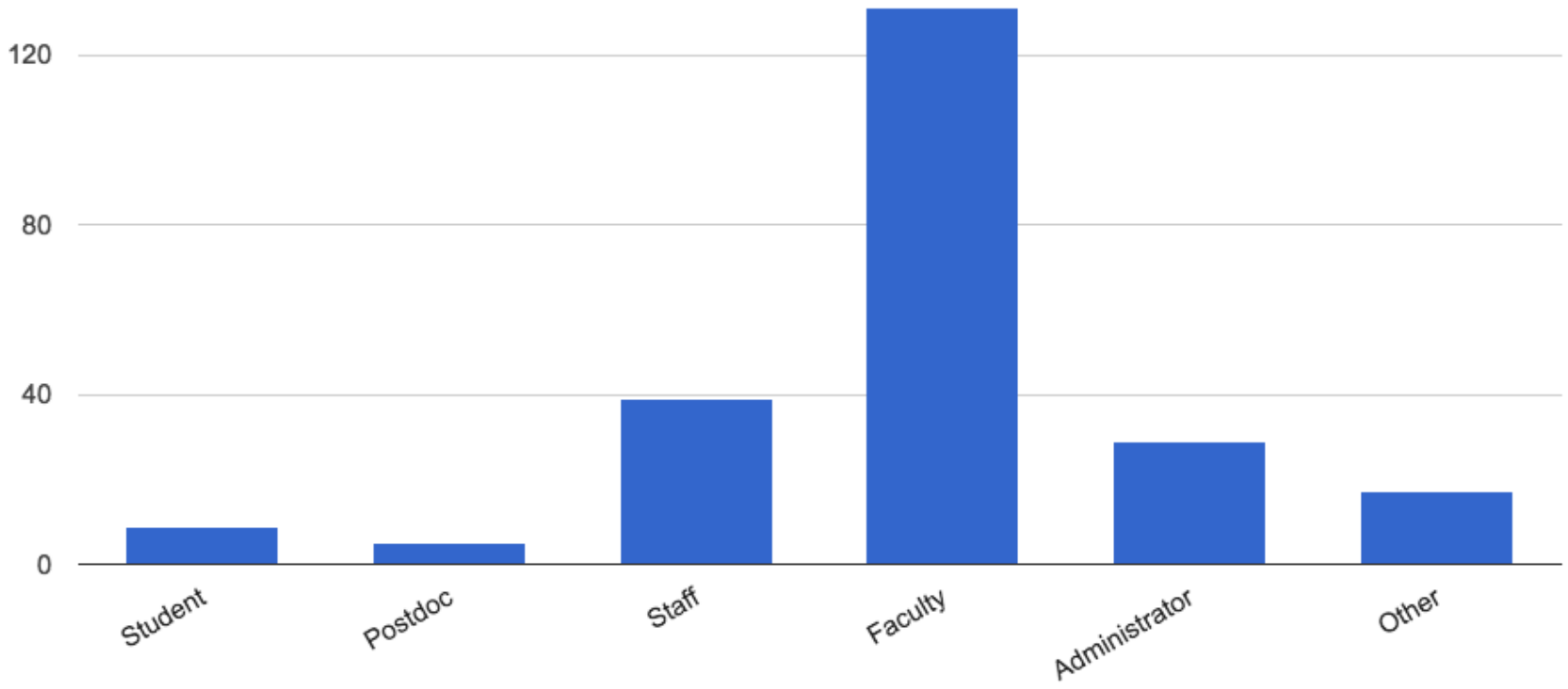
Researchers share data within research teams and labs

Researchers share data after article publication or grant completion

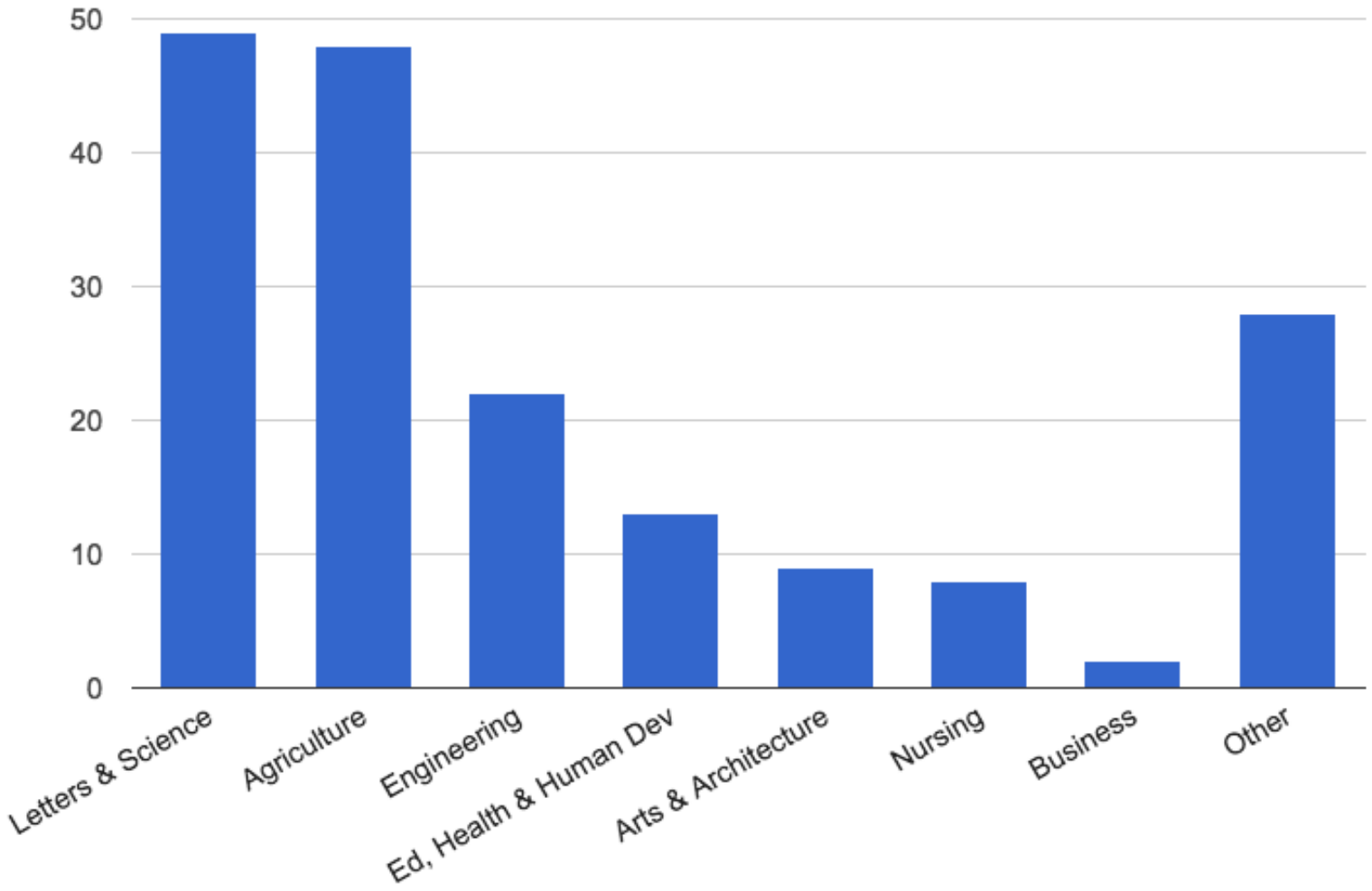
Researchers with bigger data are more likely to need campus data services

Survey Results

Demographics



Demographics



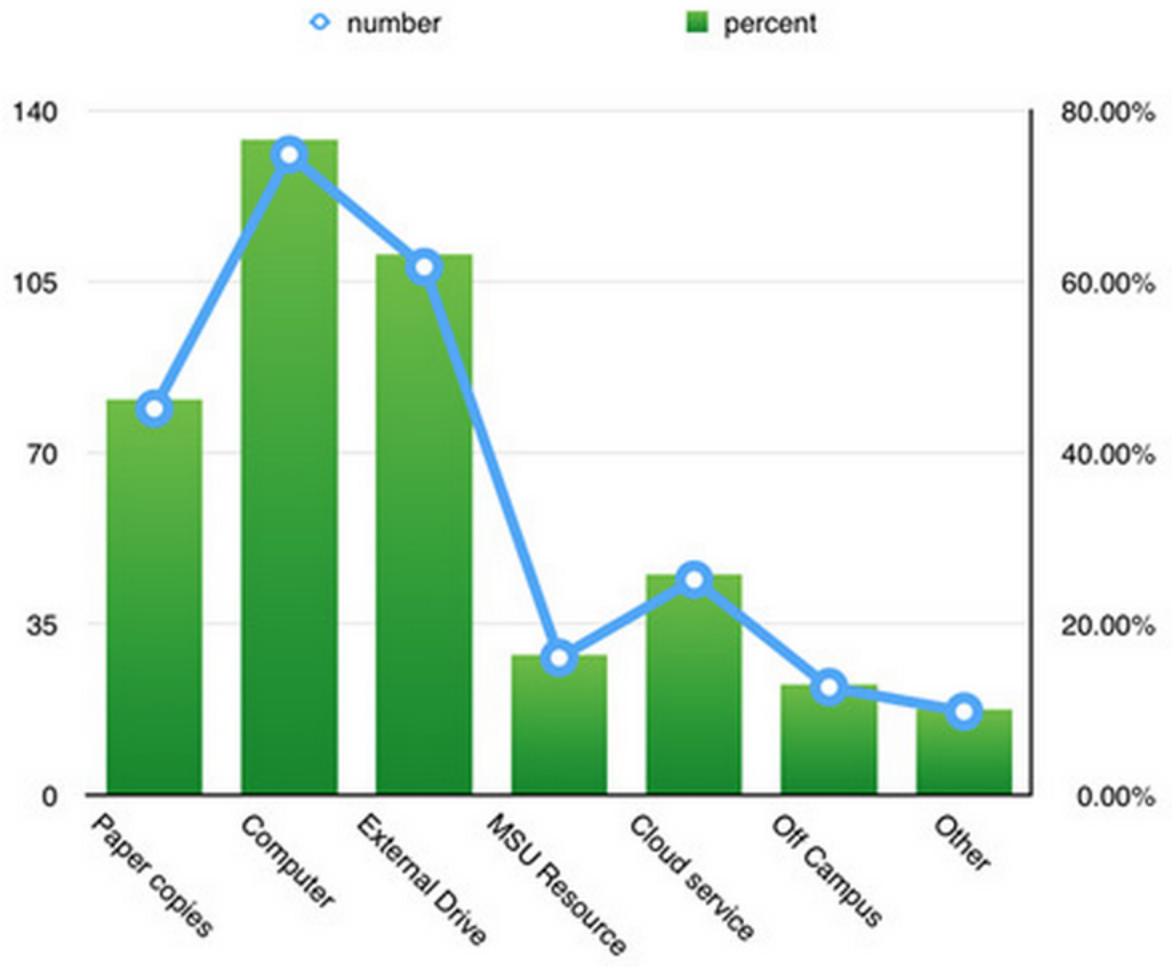
Survey Results:
Data Storage, Format, and Size

Assumptions

Most researchers store their data ~~digitally,~~ on a personal computer or external hard drive

Data Storage

	number	percent
Paper copies	79	46.20%
Computer	131	76.61%
External Drive	108	63.16%
MSU Resource	28	16.37%
Cloud service	44	25.73%
Off Campus	22	12.87%
Other	17	9.94%

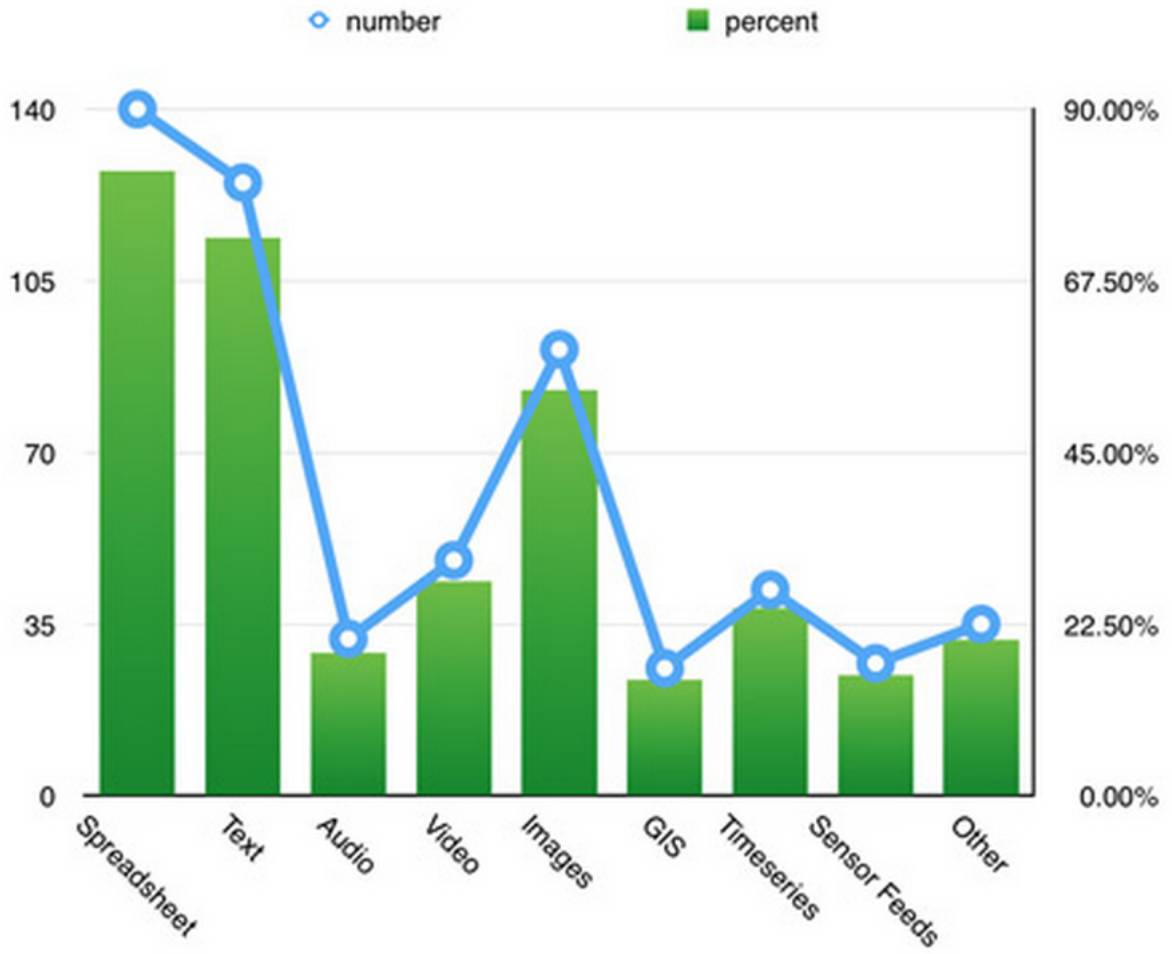


Assumptions

Researchers create ~~a wide variety of~~ data formats

Data Format

	number	percent
Spreadsheet	140	81.87%
Text	125	73.10%
Audio	32	18.71%
Video	48	28.07%
Images	91	53.22%
GIS	26	15.20%
Timeseries	42	24.56%
Sensor Feeds	27	15.79%
Other	35	20.47%

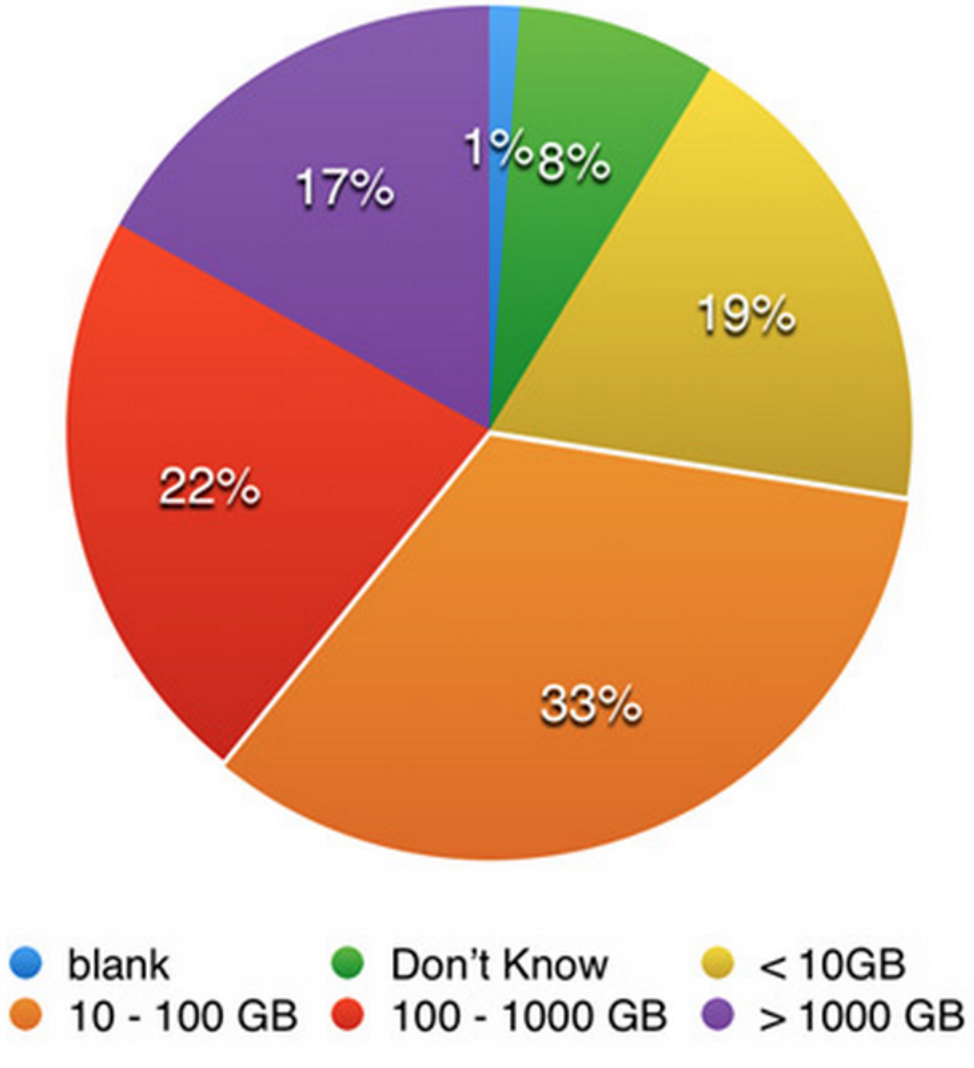


Assumptions

~~Researchers have relatively small amounts of data—less than 100 Gb~~

Data Size

blank	2
Don't Know	13
< 10GB	32
10 - 100 GB	57
100 - 1000 GB	38
> 1000 GB	29
TOTAL	171



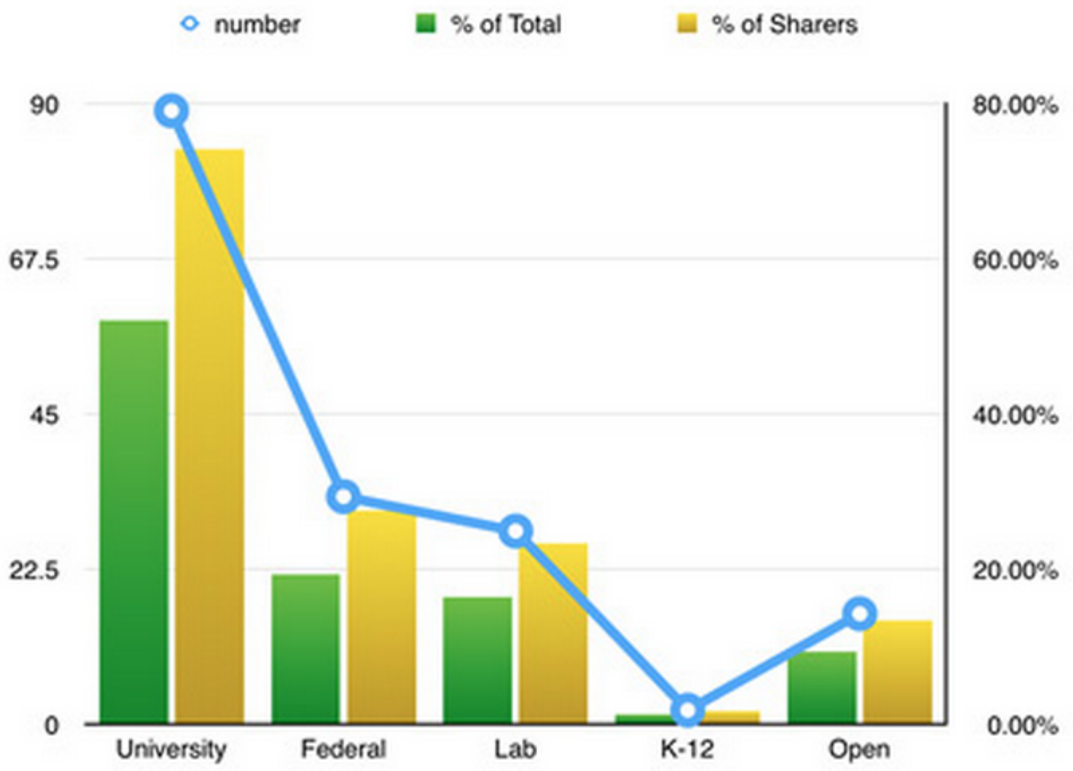
Survey Results:
Data Sharing

Assumptions

Researchers share data within research teams and labs

Data Sharing

	number	% of Total	% of Sharers
University	89	52.05%	74.17%
Federal	33	19.30%	27.50%
Lab	28	16.37%	23.33%
K-12	2	1.17%	1.67%
Open	16	9.36%	13.33%

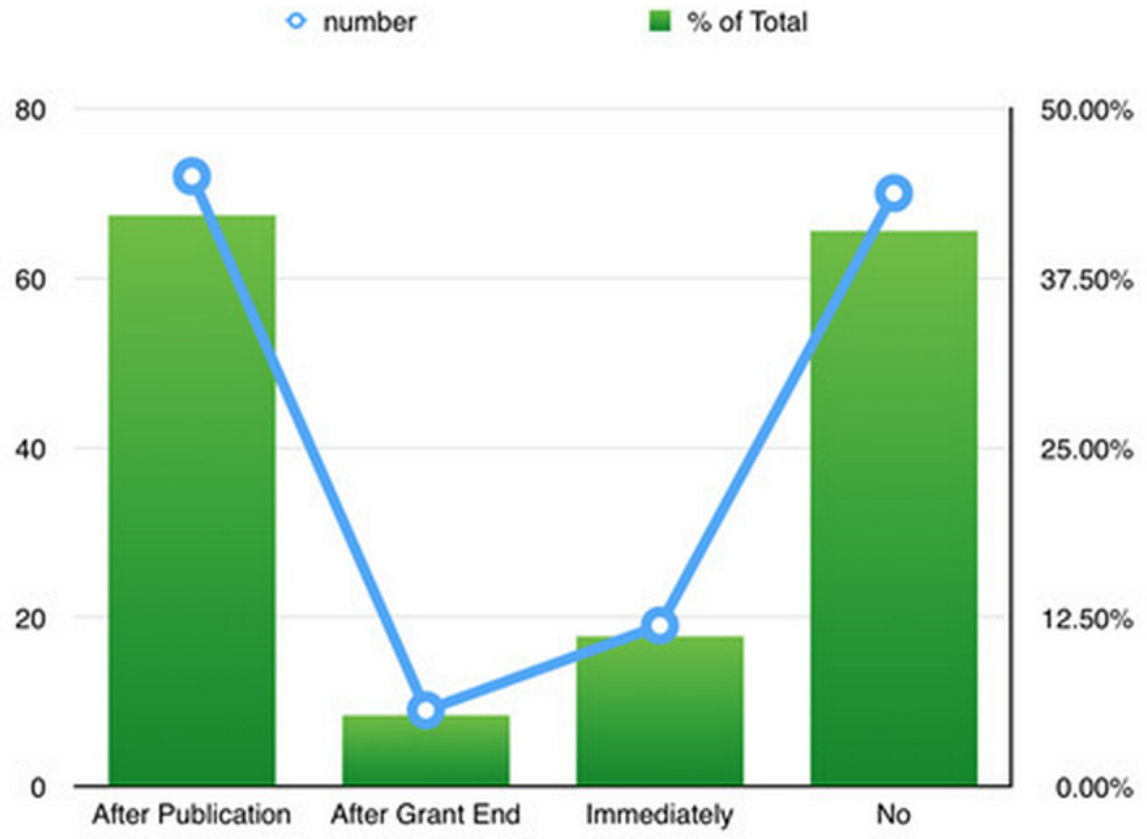


Assumptions

Researchers share data after article publication ~~or upon grant completion~~

Public Data Sharing

	number	% of Total
After Publication	72	42.11%
After Grant End	9	5.26%
Immediately	19	11.11%
No	70	40.94%



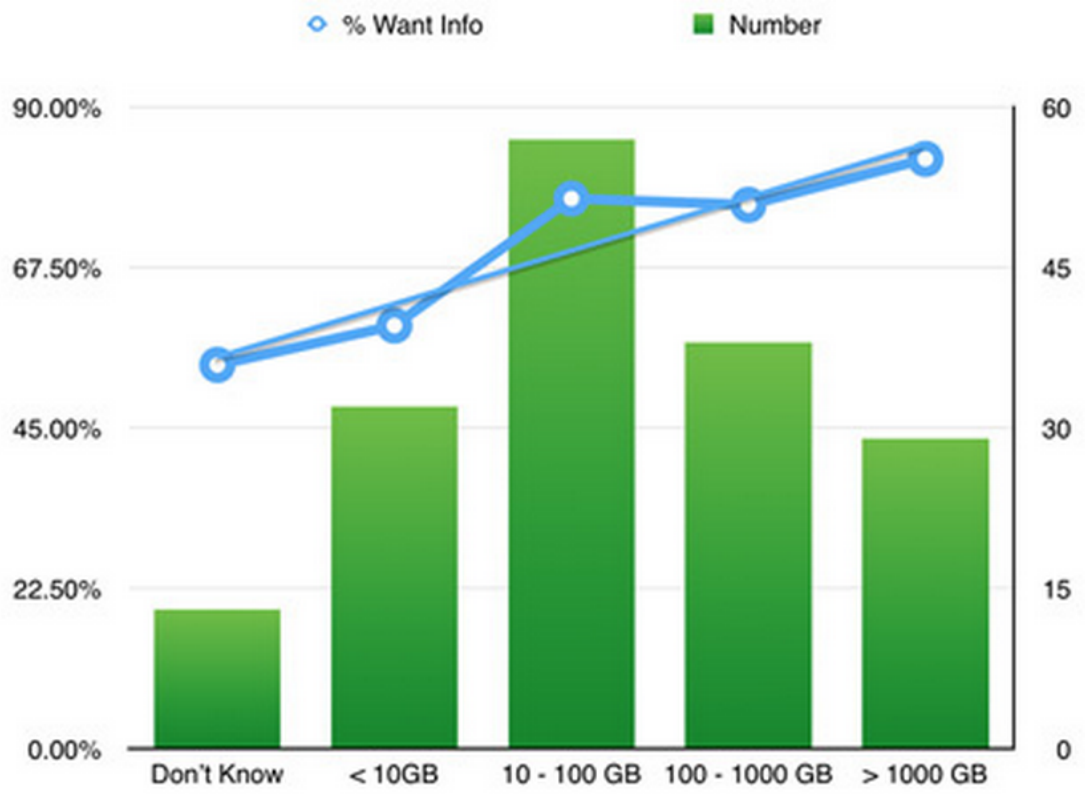
Survey Results:
Interest in Data Services

Assumptions

Researchers with bigger data are more likely to need campus data services

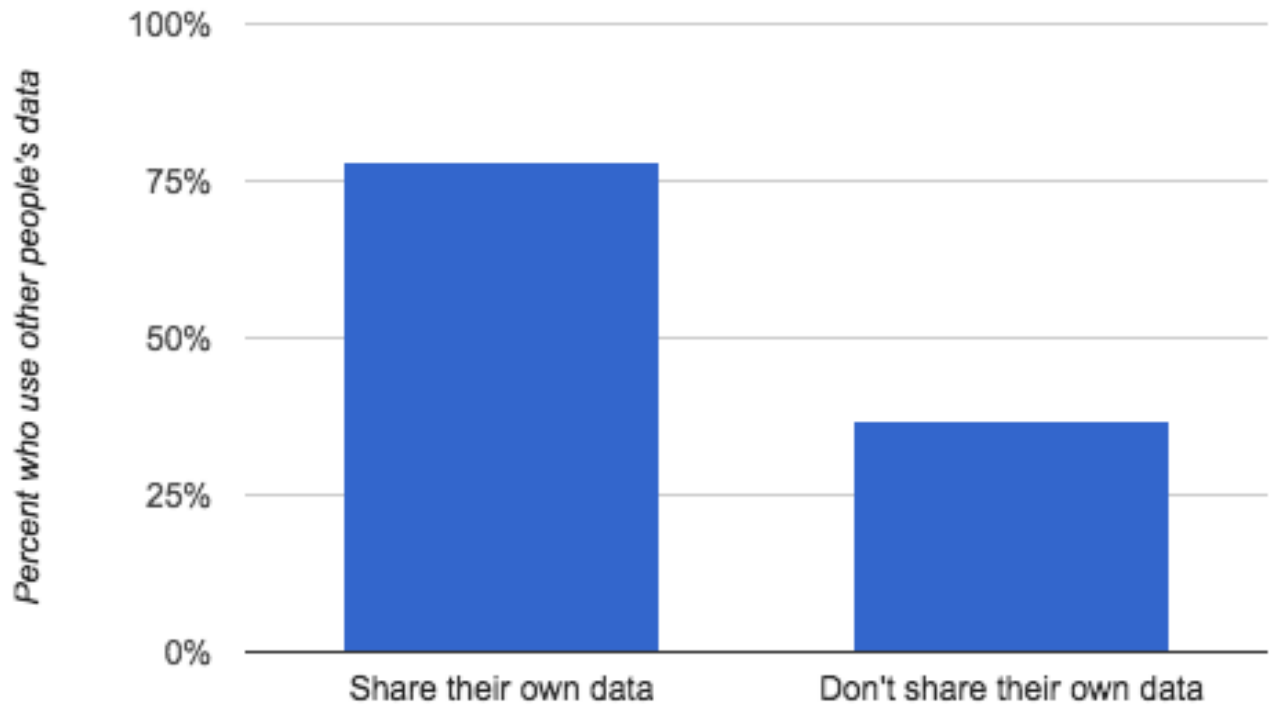
Interest in Data Services

	Number	Want Info	% Want Info
Don't Know	13	7	53.85%
< 10GB	32	19	59.38%
10 - 100 GB	57	44	77.19%
100 - 1000 GB	38	29	76.32%
> 1000 GB	29	24	82.76%



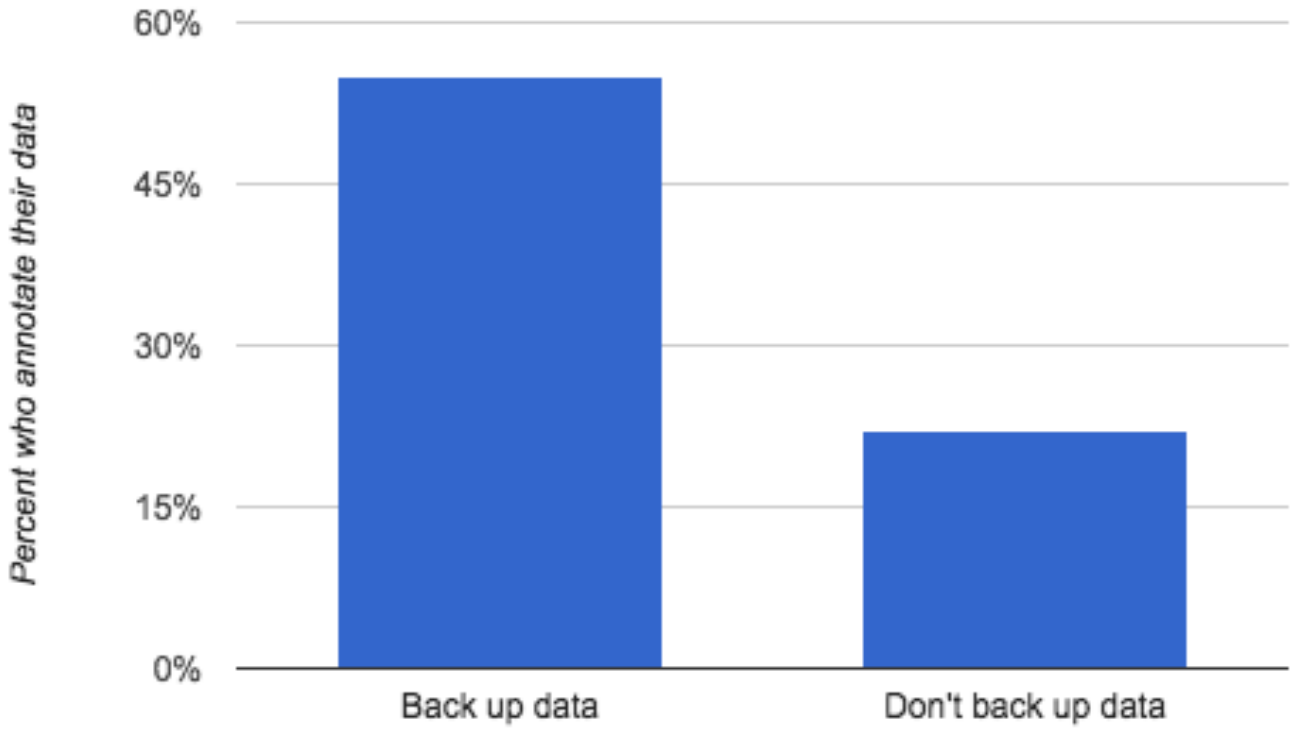
Survey Results:
Correlated Responses

Those who share data tend to use other people's data

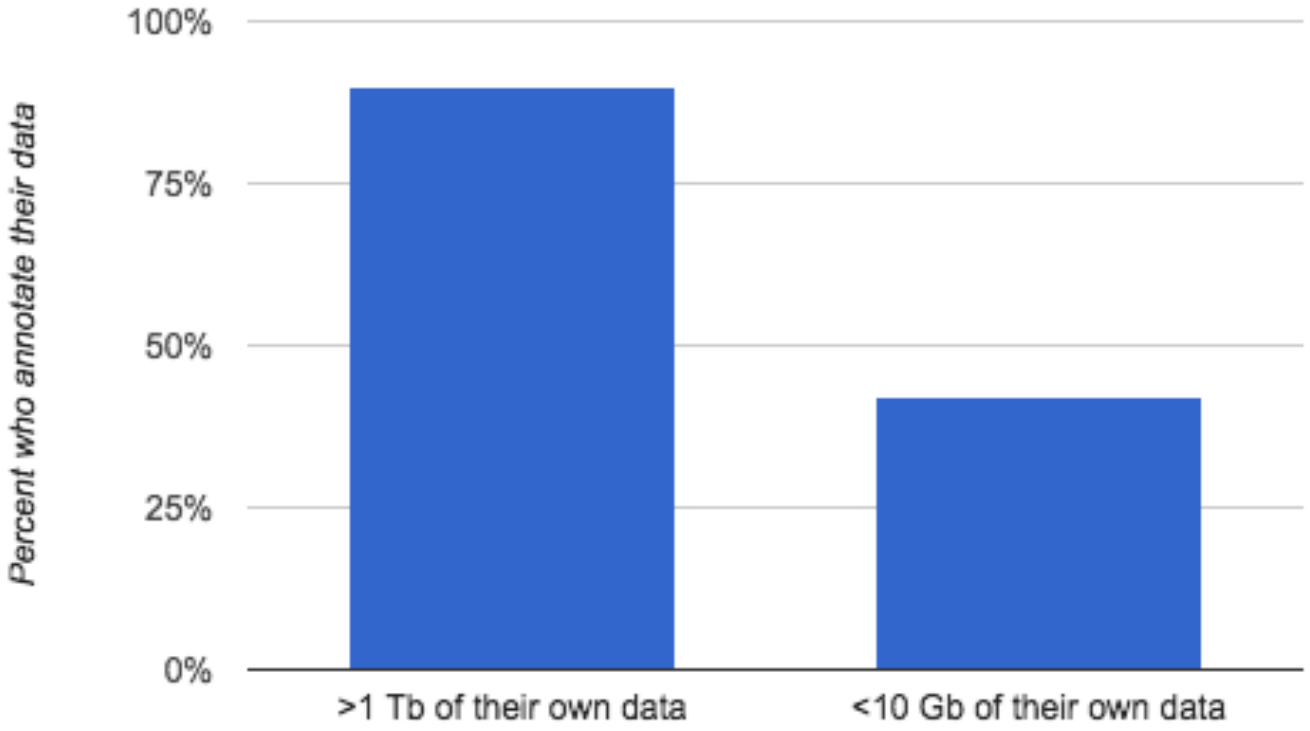


Those with larger amounts of data tend to use other people's data

Those who back up their data also annotate their data



Those with large datasets are more likely to annotate their data



Follow-up Interviews

Follow-up interviews

Interviews were scheduled with researchers from a cross-section of departments and research centers:

Electrical & Computer Engineering
Plant Sciences & Plant Pathology
Agricultural Economics & Economics
History
Land Resources & Environmental Science
Ecology
College of Engineering
Physics
Spatial Sciences Center

Interview Results:
Data Sharing

Data Sharing

Researchers weren't always sure how large their data was in terms of file size, but could describe the tools and methods they used to work with the data

Can the data be shared through e-mail attachments?
Stored on external hard drives?
Require deposit in institutional clusters or repositories?

Data Sharing

"It's important for data to be public. There should be no black boxes. We need systems that promote transparency and reproducibility in science."

"The journals where I typically publish have no data archiving requirements, so that's not [an activity] I end up doing."

"Journals with a high impact factor [in my field] require data."

Interview Results:
“Big” Data

“Big” Data

History faculty member: a 50 Kb dataset is “big” data

Another researcher working with 10 Gb of data: not “big” data, since it is smaller than previous projects.

"Managing multiple 'tiny data' sets is just as labor-intensive as [working with] big data."

Interview Results: Collaboration

Collaboration

All interviewees described challenges when collaborating and working with collaborative data across institutions:

different IT infrastructures

different data handling routines

different storage and access requirements

interdisciplinary cultural practices

Collaboration

"What we need to create is an interdisciplinary 'Rosetta Stone' to make data shareable"

"There is no language to talk about data processes that facilitate sharing"

Interview Results:
Library Love

Library Love

All interviewees said they would use MSU Library data services:

Dataset annotation and metadata markup

Assistance with deposit in relevant data repositories

Educational programs and training on campus IT resources

Data outreach and publicity

Discoveries

Discoveries

The data census created a foundation of knowledge about data on campus

Discoveries

The census will help us improve our understanding of the local research data environment

Discoveries

The census will inform Campus IT and Library priorities

Discoveries

The census confirmed that Campus IT and the Library must work closely together to meet data needs at MSU

Lessons Learned

Research other institutional data surveys and assessment frameworks

Balance length of survey with depth of information gathered

Advocate for the information the library wants to collect.

Seek IRB approval to enable data sharing

(Survey instrument: http://doi.org/10.15788/m2h59m)

Next steps

Use the data census to validate other research (e.g. Sentiment Analysis project)

Revise and redistribute the survey in early 2016

Resources

1. Averkamp S, Gu X, Rogers B (2014) Data Management at the University of Iowa: A University Libraries Report on Campus Research Data Needs. http://ir.uiowa.edu/lib_pubs/153
2. Data Asset Framework http://www.data-audit.eu/
3. Parsons T, Grimshaw S, Williamson, L (2013) Research Data Management Survey. University of Nottingham. http://eprints.nottingham.ac.uk/id/eprint/1893
4. Provost's Task Force on the Stewardship of Digital Research Data (2012) Research Data Stewardship at UNC: Recommendations for Scholarly Practice and Leadership. http://sils.unc.edu/sites/default/files/general/research/UNC_Research_Data_Stewardship_Report.pdf
5. Rolando L, Doty C, Hagenmaier W, Valk A,Parham SW (2013) Institutional readiness for data stewardship: Findings and recommendations from the research data assessment. http://hdl.handle.net/1853/48188
6. Sheehan J, Arlitsch K, Mannheimer S, Knobel C, Llovet P (2015) Data-Intensive Science and Campus IT. EDUCAUSE Review. http://er.educause.edu/articles/2015/9/data-intensive-science-and-campus-it
7. Arlistch K, Clark JA, Hager B, Heetderks T, Llovet P, Mannheimer S, Mazurie A, Sheehan J, Sterman L (2015) Montana State University Research Data Census Instrument, Version 1. MSU ScholarWorks. http://doi.org/10.15788/m2h59m
8. University of Leeds Research Data Survey https://library.leeds.ac.uk/info/377/roadmap/122/roadmap_project_outputs/7