Data-Intensive Science and Campus IT

Authors: Jerry Sheehan, Kenning Arlitsch, Sara Mannheimer, Cory Knobel, Pol Llovet

This is a postprint of an article that originally appeared in EDUCAUSE Review, September 2015.
http://er.educause.edu/

Suggested Citation: Sheehan J, Arlitsch K, Mannheimer S, Knobel C, Llovet P. 2015. Data-Intensive Science and Campus IT. EDUCAUSE Review.

Data-Intensive Science and Campus IT

Jerry Sheehan, Kenning Arlitsch, Sara Mannheimer, Cory Knobel, and Pol Llovet

Monday, September 28, 2015

Montana State University developed the Research Data Census to engage local research communities in dialogue about their data: size, sharing resources and behaviors, and interest in services. The census confirmed the need for a tight coupling of IT infrastructure to data and curation services in order to make those resources useful to the research community.

The era of data-intensive science has arrived on university campuses, with profound new demands for information technology infrastructure and services. University decision makers increasingly are asked to make major new investments in computational facilities to generate data, expand digital research repositories, and provide bleeding-edge networks.

But what do we actually know about the data on our campuses? No automated tools exist to help us discover the research silos that exist across our campuses. Driven by

economics, individual researchers often procure terabytes worth of cheap storage on institutional credit cards, leading to a haphazard array of storage, backup, security, and sharing practices. The commoditization of scientific instruments likewise has reduced acquisition costs for major data producing tools. Faculty startup packages and research grants facilitate these behaviors, producing volumes of data that just five years ago would have come from a small number of high-priced core research facilities.

Montana State University developed the Research Data Census (RDC) to engage local research communities in dialogue about their data. The RDC team was particularly interested in learning more about the following issues at Montana State:

- The size of research data
- The role that local- and wide-area networks play in accessing and sharing resources
- Data sharing behaviors
- Interest in existing services for curation, storage, and publication of scientific data discoveries

This article describes the methods, findings, and impact of the data census.[1]

About Montana State University

Montana State University is the state's public land grant university, with over 15,000 students and 4,000 faculty and staff. Montana State is listed by the **Carnegie Classification of Institutions of Higher Education** as one of 108 Research Universities of Very High Intensity (RU/VH), with annual externally sponsored research expenditures of over $100 million. Primary federal funding agencies for Montana State include the National Institutes of Health, the National Science Foundation, and the Department of Agriculture.

The Research Data Census

The RDC team consisted of members from the Information Technology Center and the MSU Library, with additional design input from the Office of Planning and Analysis. The team developed a survey instrument that provides insight into the nature of research data on campus. Qualitative data from interviews conducted with a subset of respondents also informed the findings.

One of the RDC's core design principles aims to gain a better technical understanding of the issues without requiring technology expertise from the respondents. Perhaps the best example of the RDC's approach to addressing this tension appears in the first census question, which asks, "How Large is Your Data?" Respondents could select from five answers: I don't know, <10 Gb, 10–100 Gb, 100–1,000 Gb, >1,000 Gb. In addition to these technical responses, each selection also included a nontechnical reference point: <10 Gb = fits on a USB drive, 10–100 Gb = fits on my computer, 100–1,000 Gb = fits on an external hard drive, and <1,000 Gb = 1 terabyte. This approach compromises between the granularity that specific technical size numbers would have provided and the potential respondent frustration at being unable to answer the question.

The CIO, dean of the library, and vice president for research co-sponsored the RDC. The cover e-mail, distributed to all faculty and staff, emphasized that its purpose was to gain insight into research data to help drive future infrastructure and service investments across the sponsoring organizations. In addition to the blanket e-mail, governing groups of all three organizations were briefed on the RDC and asked to encourage their members and colleagues to participate. The sponsoring organizations also committed to distributing the survey results back to campus via those same governing organizations and public forums focused on research infrastructure needs.

Key Findings from the Census

Here we present our major findings organized by the RDC's questions.

Participant Demographics

The census collected demographic information on respondents' college, gender, and role (faculty, staff, postdoctoral students, and undergraduate students), as shown in table 1.

Table 1. Demographics of Research Respondents

College	Male/Female	Faculty	Staff	Post-Doc	Student	Total
Letters & Science	26/23	38	6	3	2	49
Agriculture	31/17	37	10	0	1	48
Engineering	12/10	16	5	0	1	22
Education, Health & Human Development	3/10	10	2	0	1	13
Arts & Architecture	3/6	7	1	0	1	9
Nursing	2/6	7	1	0	0	8
Business	0/2	2	0	0	0	2
Other	14/14	8	19	0	1	28
Total	**91/89**	**125**	**44**	**3**	**7**	**179**

Where Is Research Data Stored?

Researchers store their data primarily on locally accessible computers (workstation or lab computers; 76 percent), with external drives coming in as the next most common (63 percent) and 46 percent of respondents still using paper to record data. Less than 20 percent of researchers used centralized and less than 15 percent used managed off-campus storage locations, with the exception of cloud services — still popular at 26 percent. See figure 1.

	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%

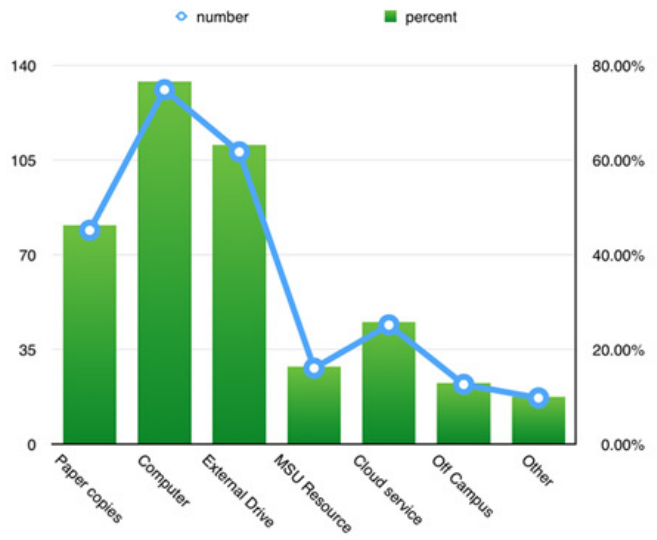

Figure 1. Where researchers store their data

What Are the Formats of Research Data?

Understanding data formats is critical for data preservation and archiving. Unsurprisingly, researchers identified spreadsheets as the dominant research data type, with nearly 82 percent storing at least some of their data in this form. Raw text data was not far behind, with 73 percent using this storage format. Images and video were the next most popular data formats (53 percent and 28 percent, respectively). The two most interesting data points are that 25 percent of the researchers have time-series data, and 20 percent store data in "Other" formats. The former indicates that time-series–based infrastructure could benefit a sizeable portion of researchers, and the latter shows that one in five researchers might require custom data management. See figure 2.

	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%

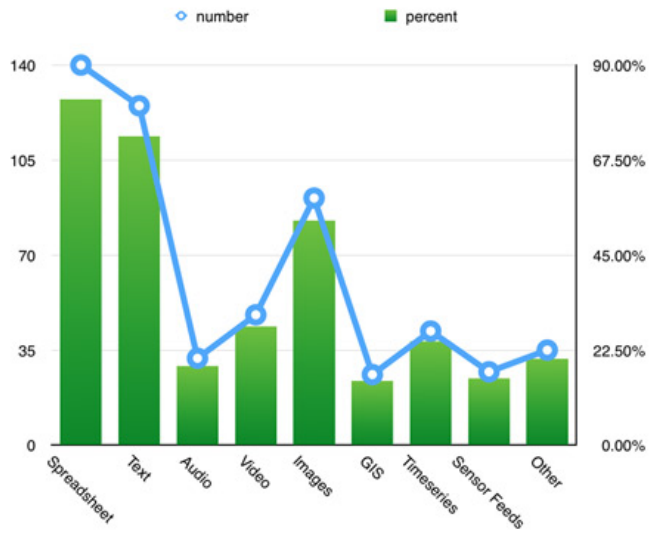

Figure 2. Data formats used

How Large Is the Data Generated by Researchers?

Size is of the utmost importance when making strategic decisions about research data storage. One-third of the researchers had data in the 10–100 Gb range, i.e., the "computer" range of storage. On either side of that, 22 percent had data between 100 and 1,000 Gb, and 19 percent had less than 10 Gb of data. Only 17 percent of researchers had data in excess of 1,000 Gb. These numbers are similar, but correspondingly larger (due to the time that has ensued), to those from a survey conducted by *Science* of its peer reviewers in 2011.[2] See figure 3.

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

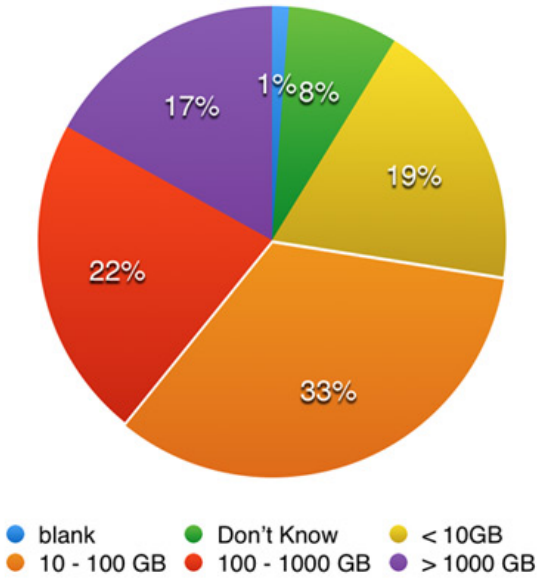

Figure 3. Size of data generated

Do You Share Your Research Data?

Ninety percent of MSU researchers said that they share their research data. Research collaboration and teaching purposes are the most likely reasons, though the RDC did not request clarification.

With Whom Do You Share Your Research Data?

Nearly 75 percent of the researchers at Montana State who share data off-campus share it with other universities. This finding is unsurprising given the collaboration and peer review of scientific research. However, given the typical sources of funding, it is surprising that fewer than 30 percent of researchers who share data share with federal agencies and research labs. Only 10 percent of researchers make their data openly available via websites and other repositories. See figure 4.

	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%

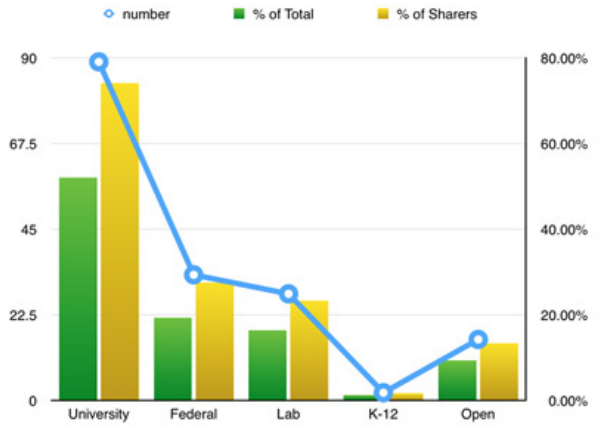

Figure 4. Recipients of shared data

When Do You Share Your Data?

The timing of data sharing (within the timeline of publication and research data gathering) is also important. Nearly 50 percent of researchers sharing data do not do so publicly, opting instead to keep their data restricted (sharing with collaborators, labs, agencies, etc.). Still, 42 percent share their data after publication, with 5 percent waiting until the end of the grant to publicly share their data. Only 11 percent of researchers choose to share their data immediately. See figure 5.

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

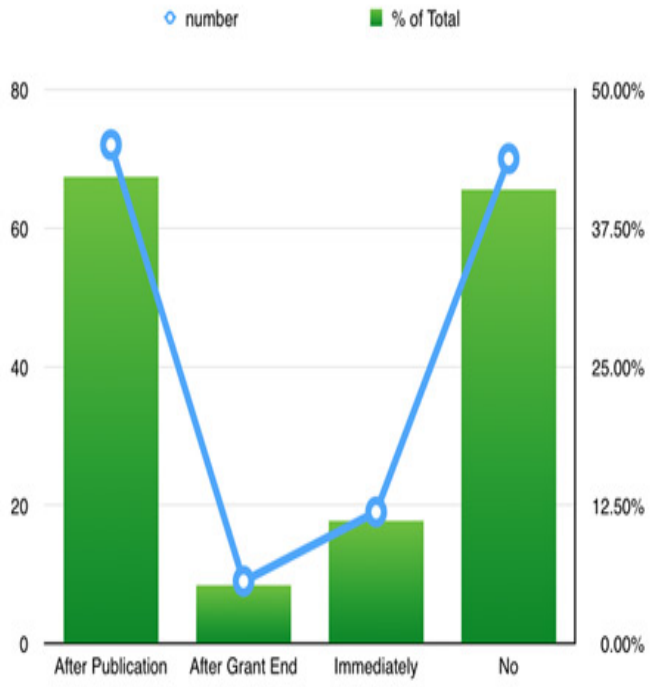

Figure 5. Timing for sharing data

Do You Have Interest in Additional Data Services?

Seventy-three percent of researchers expressed interest in learning more about data services and infrastructure. We hypothesized that data size would correlate with the desire for additional service information, expecting that researchers having more data will have more interest in managing their data. This hypothesis is supported by RDC responses. Of respondents who reported not knowing the size of their data, only 54 percent expressed interest in more information about data services; of respondents who reported having >1,000 Gb of data, 83 percent expressed interest in more information. See figure 6.

	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%

Figure 6. Research interest in data services

Analysis of Findings from the Research Data Census

A number of statistically significant findings surfaced from the quantitative component of the RDC:

- Researchers who share their data, regardless of who they share it with (colleagues, students, or non-MSU researchers) also tend to download data from other sources or repositories (78 percent of people sharing their data also download data, versus 37 percent of

people not sharing their data; *p*-value: 1.67×10^{-7}).

- Researchers with large research data tend to download data from other sources or repositories (90 percent of people with data sets above one terabyte also download data, versus 42 percent for people with data sets below 10 Gb; *p*-value: 1.58×10^{-5}).
- Researchers who back up their data also tend to annotate it (55 percent of people who back up their data also annotate it, versus 22 percent of people who don't back up their data; *p*-value: 5×10^{-3}).
- Researchers with large research data tend to annotate it (62 percent of people with data sets above one terabyte also annotate their data, versus 39 percent of people with data sets below 10 Gb; *p*-value: 0.024).
- Researchers interested in learning more about data infrastructure and services who do not back up their data cite technical barriers as their main reason for not doing so (*p*-value: 0.014).

Qualitative Interviews of Researchers

Survey respondents were asked if they would be available for follow-up questions, and eight individuals were selected through purposive sampling for a one-hour qualitative interview to explore data practices, perceptions, and usage in more depth.

Insights from the RDC qualitative interviews follow:

- While researchers could roughly estimate the amount/size of data that they generate, analyze, and store, their connection to data does not revolve around file sizes, but rather data transfer and the tools used to move and work with their data. For example, researchers more naturally described

whether data can be shared through e-mail attachments, placed on external hard drives, or requires deposition in institutional clusters or repositories.

- Researchers' sense of when and how data is disseminated and shared varied widely. A minority is dedicated to open access and sharing practices throughout the research process, with statements such as, "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 remainder advocate for sharing finalized data sets at the conclusion of a study or concurrent with publication. Still, there was also a sense that external requirements for data sharing provide incentive for the work involved, as one researcher remarked: "The journals where I typically publish have no data archiving requirements, so that's not [an activity] I end up doing." Another claimed that such data dissemination is *de rigueur*: "Journals with a high impact factor [in my field] require data."
- There was no consensus regarding the definition of "big data," and the fuzzy boundaries of this term might introduce both false positives and false negatives in research questioning. For example, a humanities faculty member remarked that within his discipline, a data set with 50K entries would be considered big data, while a researcher who was working with a data set of several gigabytes did not consider the project to be big data, since it was smaller than the typical data set size found in previous projects. Further, the overall size of data for a researcher might be based on an aggregation of many resources, as one researcher observed: "Managing multiple 'tiny data' sets is just as intensive as [working with] big data." To uncover the nuances of data practices among researchers, the use of more

naturalistic terms might be beneficial.

- Without exception, interviewees described their research practices as involving collaboration with others, both inside and outside the institution. These collaborative practices require negotiation of heterogeneous IT infrastructures, data handling and modification routines, storage and access technical requirements, interdisciplinary cultural practices, and metadata and annotation standards. Frequently, though, researchers remarked that collaboration across boundaries — both disciplinary and institutional — was as much an ontological issue as a technological one, with respondents remarking, "What we need to create is an interdisciplinary 'Rosetta Stone' to make data shareable" and "There is no language to talk about data processes that facilitate sharing."
- All researchers responded positively when asked if they would engage MSU Library services that focus on data set annotation and metadata markup, assistance with deposit in relevant data repositories, and educational programs and training on campus IT resources. Several interviewees made suggestions for collaboration with the library, such as a regular e-mail newsletter that lists newly submitted data sets from campus researchers that are available for use.

By pairing qualitative interviews with the results of the RDC, we gained insight into many of the contextual factors that drive responses to survey questions. First, our follow-up interviews validated the RDC's strategy of emphasizing practical definitions; researchers described their data in terms of storage methods, and they defined "big data" relative to their specific discipline and experience. By framing the complexities of research data management with real-world storage methods and routines, we bridge gaps in understanding caused by different jargon and varied

technical expertise.

Second, both the RDC and the follow-up interviews showed that researchers' attitudes toward open access and data sharing vary widely.

Third, as researchers discussed collaborative efforts across disciplines, campuses, and countries, these interviews revealed a need for research cyberinfrastructure plans to recognize that "the campus" is a potentially artificial boundary when describing the scope and governance of IT resources and data practices.

Impacts of the RDC

Montana State created the Research Data Census to improve our understanding of the local research data environment. The most profound impact of the RDC has been a fundamental change in the culture that drives IT infrastructure and services investments. Previously, capital investments affecting networking were made on the basis of lifecycle replacement (age of the switching infrastructure) or in a "keeping up with the Jones's mentality" (the campus needs a 100 Gbps wide-area network to match other R1 universities). The insight generated by the census has allowed Montana State to move beyond this mentality and toward data-driven investment decisions.

The RDC also significantly affected engagement on campus because it solicited broad user input on the greatest opportunities for capital investments. These investments promise to increase our research competitiveness and scientific discoveries, and we can demonstrate that campus researcher input drove the allocation of FY16 capital.

The census confirmed the need for a tight coupling of IT infrastructure to data and curation services in order to make those resources useful to the research community. As a result, a new partnership between the Information

Technology Center and the MSU Library will build a collective that synthesizes infrastructure, personnel, and services to address the continuum of need, from data production to storage, curation, discovery, and long-term preservation.

The Montana State University system has adopted the RDC beyond the Bozeman campus as a useful tool for informing research infrastructure investments at scale. Based on results from the study described here, we have refined the survey instrument and believe that wide availability can provide valuable benchmarking to the wider EDUCAUSE community. To this end, Montana State will be releasing an updated community version of the Research Data Census in late 2015 with the aim of facilitating regional and national data collection and analysis on data practices and requirements for research institutions.

Research Data Census Development Team

- Kenning Arlitsch, Dean of the Library
- Jason A. Clark, Head of Library Informatics and Computing
- Ben Hager, Library Systems Administrator
- Thomas Heetderks, Software Engineer
- Pol Llovet, Associate Director of Research Cyberinfrastructure
- Sara Mannheimer, Data Management Librarian
- Aurelien Mazurie, Director of Bioinformatics Core
- Jerry Sheehan, Chief Information Officer
- Leila Sterman, Scholarly Communication Librarian

Notes

1. The **survey instrument** is available in Montana State University ScholarWorks; DOI: 10.15788/m2h59m.
2. *Science* Staff, "**Challenges and Opportunities**," *Science*, Vol. 331, No. 6018 (February 11, 2011): 692–693; DOI: 10.1126/science.331.6018.692.

Jerry Sheehan is the chief information officer for Montana State, where he has responsibility for the infrastructure that supports the learning, discovery, and outreach missions of the university. Prior to his arrival in 2014 at Montana State, he was the chief of staff for the California Institute for Telecommunications and Information Technology (Calit2) at UC San Diego. Sheehan has also served as an associate vice president for Information Technology at Purdue University, a deputy director for the National Center for Supercomputing Applications, and an assistant to the Lt. Governor of Illinois.

Kenning Arlitsch is dean of the Library at Montana State University, where he leads a progressive research library that supports student success, the university's research enterprise, and statewide cooperative efforts. Prior to moving to Montana he was associate dean for IT Services at the J. Willard Marriott Library at the University of Utah. He is the founder of several notable digital library programs, including the Mountain West Digital Library and the Utah Digital Newspapers, and he is co-founder of the Acoustic Atlas. His funded research for the past five years has focused on search engine and semantic web optimization for digital repositories.

Sara Mannheimer is data management librarian at Montana State University. After receiving a Master's in Information Science at the University of North Carolina at Chapel Hill, she

worked as curator of Dryad Repository, where she led the effort to draft the repository's data preservation policy. At Montana State University, she continues to focus her work and research around data preservation, data management, and open data advocacy.

Cory Knobel is CEO of Research At Work Consulting. Formerly a professor in Information and Computer Science at the University of California–Irvine and University of Pittsburgh, he now provides strategy and management consulting to data-intensive organizations. Knobel received his PhD in Information from the University of Michigan School of Information and focused his research career on the coordination of interdisciplinary work and the organizational dynamics and evaluation of scientific research cyberinfrastructure.

Pol Llovet supports research at Montana State University as associate director of research cyberinfrastructure and technical director of the Hyalite Community HPC Cluster, Science DMZ, and research storage. He is a cyberinfrastructure generalist with expertise in software engineering, high-performance computing, parallel file systems, devops, and research data management.