Data Management Practices and Needs at the University of Rochester

Kathleen Fear
December 13, 2013

Executive Summary

Key points:

- Humanists responded relatively strongly: it isn't just scientists who are engaged with issues around data management.
- Researchers are interested in sharing their data. Very few said none of their data could be shared for any reason, and very few expressed interest in limiting their sharing only to collaborators.
- The service respondents expressed the most interest in is an institutional data repository.
- The services respondents are interested in align with what we're working on already—but survey responses indicate room to improve.

This survey documents the diversity of data practices at the University. Researchers across fields work with data, much of it digital, and a majority need to document their data management practices for their funders. Many are engaged in producing new data, but a substantial number also work with secondary data and simulations. While numeric data makes up the bulk of the data with which they work, researchers at the University also handle a variety of other kinds of data, some of it very complex; they store their data in a range of different forms and formats, some of which will pose significant preservation concerns.

While scholars in the humanities were more likely to indicate that they do not work with data than researchers in other fields, humanists who do use data were a substantial proportion of the respondents to this survey. Given the increasing investment in and visibility of the digital humanities at the University, it will be important to consider the distinct needs of humanists when designing data policies and services.

Researchers are, by and large, interested in sharing some or all of their data in the future. While many scientists are already making use of disciplinary repositories, researchers in other fields may not have found ways to share their data yet. Respondents reported that much of their data could be shared without restriction; where data couldn't be shared, ethical and privacy concerns were a primary barrier.

While this diversity of practices and needs may present a challenge to the design of institutional policies and services, there were several common needs across respondents and disciplines. Two services emerged as especially of interest to

researchers here: the development of an institutional data repository, and support in creating data management plans. Other services, including help finding and using disciplinary repositories, were popular among particular disciplines, but across the board, these two services received the strongest support. These preferences align with efforts that are already ongoing in the University and the Library: the RCL has developed several different approaches to supporting data management planning, and there are several options for researchers to store their data on campus.

However, researchers' interest in these services suggests that there is room for improvement. The Libraries' data management support services are fairly new, and many faculty may not yet know about them. And there is no central location for research data storage that can accept large datasets on campus: the institutional repository managed by the Libraries, UR Research, makes data publicly accessible, but it does not (and will not be able to) accept large datasets, whereas storage with the Primary Data Center or in departments may handle larger datasets but does not have a public-facing access point.

Introduction

Researchers at the University of Rochester produce increasingly massive amounts of data, and funding agencies are taking an increasing interest in the management and preservation of that data. This survey of faculty in Arts, Sciences and Engineering was undertaken in order to gain a better understanding of the landscape of data practices at the University as well as to examine opportunities for introducing new services to support data management and improving existing ones. The survey explores what data these researchers are working with; how they currently manage their data; and what their interests are for services that the University could provide.

Current support for data management

Support for data management encompasses infrastructural concerns, such as storage space, network capacity and security, as well as education, outreach and consultation. At the University of Rochester, we have made some strides in these areas.

The University of Rochester's River Campus Libraries have been actively building capacity in data curation and management. A particular area of focus has been support for faculty creating data management plans for funders, including offering the customized DMPTool, the Data Management website, workshops on data management planning, and consultations with the data librarian.

There are also options for data storage and preservation on campus. The Warner School, for example, offers its faculty members storage through a contract with the University's Primary Data Center. Within Arts, Sciences & Engineering, storage is managed either locally by departmental systems administrators, or centrally through University IT. The Libraries also offer the UR Research repository. However, that repository can only take small deposits, on the order of a few gigabytes, whereas many researchers, especially in the sciences, may need storage for multiple terabytes of data.

These activities are occurring within the broader context of an increasing emphasis on Big Data on campus and the beginning of the University's data science initiative, a $100 million commitment to build an interdisciplinary Data Science Institute. Specific to data curation and digital scholarship in the humanities, the University was recently awarded a $1 million Mellon grant to fund a tenure-track faculty position in digital media and several graduate student fellowships in digital humanities. Further, our institution is part of the Mellon Central New York Humanities Corridor, which energizes digital humanities work in our region. President Joel Seligman recently announced an additional $100,000 funding for the humanities at Rochester, raising special humanities funding to $250,000 annually. The University's commitment to data science across disciplines underscores the importance of campus-wide policies and services that will support researchers in producing and managing their data: it is important not only to support researchers

in their cutting-edge work but to ensure that the data that underlies that work is protected and maintained over the long term.

Methods

Several other universities have conducted similar assessments of their data management practices and needs, including the University of Michigan, the University of California at San Diego, Oxford University, Cornell University and a consortium of Australian colleges and universities. The University of Rochester survey was developed primarily by the data librarian based on these existing surveys, especially the University of Oxford's data management assessment. Prior to releasing the survey, RCL's subject librarians along with Rob Clark and David Williams reviewed the questions and provided feedback.

David Williams' office provided names and email addresses for 343 faculty in the Schools of Art, Sciences and Engineering. The survey was released on Monday, October 28 and remained open until Friday, November 8.

Findings

343 faculty members received the survey through SurveyMonkey at their University email addresses. Of the 343 survey recipients, 67 provided complete responses, for an overall response rate of 20%. All respondents were affiliated with the Schools of Art, Science & Engineering. One was also affiliated with the School of Medicine and Dentistry. No respondents were affiliated with Warner, Simon, the School of Nursing, or Eastman.

Ten respondents indicated that they do not produce or work with data in their research; their responses are excluded from this analysis. These individuals were all in Humanities fields: four in Modern Languages & Cultures; two in Art & Art History; two in English and one each in Philosophy and Religion & Classics. The 57 respondents who work with data represented 25 departments. The largest proportion of respondents came from Chemistry, Physics and Astronomy and Mechanical Engineering, with seven responses from each department. Note that there are more total affiliations than respondents, since some respondents were associated with more than one department.

Department	Number of respondents	Percent of total respondents
Sciences		
Chemistry	7	11%
Physics & Astronomy	7	11%
Brain & Cognitive Science	4	6%
Mathematics	3	5%
Biology	3	5%
Optics	3	5%
Earth & Environmental Sciences	2	3%
Computer Science	1	2%
Engineering		
Mechanical Engineering	7	11%
Electrical & Computer Engineering	4	6%
Biomedical Engineering	2	3%
Chemical Engineering	1	2%
Social Sciences		
Clinical & Social Psychology	3	5%
Political Science	3	5%
Anthropology	2	3%
Economics	1	2%
Sociology	1	2%
Humanities		
English	3	5%
History	3	5%
Linguistics	1	2%
Modern Languages & Cultures	1	2%
Religion & Classics	1	2%

Most responses came from faculty in the sciences and in engineering disciplines. The social sciences and humanities both represented 16% of the respondents with data. This distribution is similar to the overall distribution of faculty at the University ($X^2(3, N = 67) = 4.22$; $p = 0.24$). Four respondents were affiliated with departments in different disciplines; they are listed here as multidisciplinary. Three of the four multidisciplinary responses were faculty with appointments in science and engineering departments, and one had an appointment in a social sciences field along with a science.

Discipline	Number of respondents	Percent of total respondents
Sciences	24	42%
Engineering	11	19%
Social Sciences	9	16%
Humanities	9	16%
Multidisciplinary	4	7%

Across the disciplines, about a quarter of the respondents stated that they conduct their research as an individual; roughly an equal amount conduct their research as a team, with each team member looking after his or her own data. Close to half (42%) conduct some of their research as part of a team and some on their own. A minority (7%) work with data that is shared and managed within a team.

In the social sciences and in engineering, the majority of the respondents (67% and 73%, respectively) indicated that their research is a mix of team-based work and individual work. Respondents in the sciences more frequently (42%) work primarily as a team, with individuals responsible for managing their own data. The humanities were the only discipline in which it was most common to work only individually, with 67% of respondents reporting that they work primarily on their own.

Do you conduct your research as part of a team or as an individual?

Answer Options	*Number of Respondents*	*Percent of Total Respondents*
Some of my research is undertaken as part of a team but I also conduct some research independently.	24	42%
As an individual.	15	26%
As part of a team but each member of the team looks after their own data.	14	25%
As part of a team with our research data managed by the team.	4	7%

In free text responses, two respondents noted that, as heads of their labs, they create policies that individual students must follow.

Overall, about half of respondents primarily create new data in their work (53%). This is especially true in the sciences and social sciences: 71% of respondents in the sciences stated that they typically create new data. The other disciplines showed a somewhat different picture. In engineering, the largest group of respondents (36%) indicated that they work with a combination of data types, producing new data as well as using data from other sources. Engineering was also the discipline with the largest number of respondents working with simulations (27%). In the social

sciences, most (56%) respondents typically produce new data, and a third (33%) work with multiple data types. The humanities were nearly equally split between those who produce new data (44%) and those who use secondary data (44%), with the remainder of the population indicating that they work with multiple data types (11%).

Which statement best describes your use of data in your research?

Answer Options	*Number of Respondents*	*Percent of Total Respondents*
I typically work with new data that I have created or collected myself.	30	53%
I typically work with multiple data types.	11	19%
I typically work with secondary data (data that another person or organization created).	9	16%
I typically work with simulations.	7	12%

Most respondents (68%) work with numeric data, followed by statistical (40%), textual (39%) and images (39%). Almost all respondents in the sciences (92%) work with numerical data; the next largest proportion in that discipline (38%) work with textual data. Similarly, large numbers of engineering respondents (82%) indicated that they work with numeric data, but nearly as large a group (64%) also work with images. The social scientists primarily responded that they work with statistical data (78%) as well as numerical data (44%). The humanities were the only group with a large proportion of respondents stating that they work with textual data (78%). About a third (33%) of humanists also work with images. Some respondents indicated that they also manage types of data that were not listed, including visualizations, spectra, and proprietary instrument data (Bruker NMR, Agilent GC, etc.).

What kinds of data do you work with? (Or, what kinds of data does your team work with, if your data are managed collectively?)

Answer Options	*Number of Respondents*	*Percent of Total Respondents*
Numerical	39	68%
Statistical	23	40%
Textual	22	39%
Images	22	39%
Bibliographic	11	19%
Audio	9	16%
Multimedia	5	9%
Geospatial	4	7%

Most commonly, respondents stored their data in tables or spreadsheets (56%) or in word processing files (51%). Respondents also indicated a wide range of

additional formats in their open ended responses, from handwritten notes and paper printouts to instrument- and discipline-specific formats. Respondents in the social sciences rely more heavily on statistical data than other fields do, while engineers and humanists reported using images more than scientists or social scientists do.

How are your data stored or structured? (NOTE: Data need not be stored digitally! If you work with data that are not in an electronic format, please include these in your answers.)

Answer Options	*Number of Respondents*	*Percent of Total Respondents*
In tables or spreadsheets	32	56%
In word processing files	29	51%
In unstructured databases	12	21%
In relational databases	9	16%

Half of the respondents said they stored their data in a mix of open and proprietary formats; about 35% store their data mostly in proprietary formats like Excel or SPSS files.

What file formats are your data stored in?

Answer Options	*Number of Respondents*	*Percent of Total Respondents*
A mix of open and proprietary formats	26	50%
Mostly in proprietary formats (e.g. Excel SPSS)	18	35%
Mostly in open formats (e.g. delimited ASCII text)	8	15%

Roughly half the respondents (48%) felt that the time they spend managing data is helpful but not very significant to their work. About 29% indicated that their work benefits from the time they spend managing data, while 23% felt that time spent managing data is a distraction from their real work. Engineering respondents tended to be slightly more negative about data management: about 27% said data management is a distraction, and none stated that their research benefits from time on data management. By contrast, the social sciences were more positive, with about two thirds of respondents indicating that their work benefits from time spent managing data.

How important is data management to your research?

Answer Options	*Number of Respondents*	*Percent of Total Respondents*
Time spent managing data is helpful but it's not a very significant aspect of my work.	25	48%
My research benefits from the time spent managing data.	15	29%
Devoting time to managing data is a distraction from the real work of research.	12	23%

The services respondents were most interested in were an institutional repository or databank for preserving and sharing research data; assistance preparing data management plans or data sharing plans for funding; and assistance with documenting data for sharing with others or saving for later use. Interest in these three services was expressed across disciplines: in all areas, majorities of respondents were somewhat or very interested in a data repository, assistance with data management plans and documentation support. Of these, the institutional data repository had the strongest level of interest, with 58% of respondents somewhat or very interested in using that service.

While assistance identifying or using appropriate disciplinary repositories was not popular overall, respondents in the social sciences and engineering expressed interest in this service. Close to two thirds of respondents in both fields said they would be somewhat or very interested in getting help finding and submitting data to external disciplinary repositories.

Please rate your level of interest in using the following services if they were offered by the University.

Answer Options	*Somewhat or very interested*	*Neutral*	*Not very or not at all interested*
Institutional repository/databank for preservation and/or sharing of research data	58%	14%	26%
Assistance preparing data management plans or data sharing plans for grant applications	52%	21%	25%
Assistance with documenting data for sharing or saving (i.e., metadata creation)	50%	16%	32%
Workshops on data management for students/technicians/administrative assistants/postdocs	45%	24%	29%
Assistance with confidentiality/ privacy/legal/P issues associated with data preservation and/or sharing	44%	12%	42%
Assistance with identifying or using appropriate disciplinary or other external data repositories/databanks	43%	22%	33%
Workshops on data management for faculty	42%	28%	28%
Digitization of print or other types of physical records, such as lab notebooks	39%	14%	45%
Personalized consultation on data management practices (for specific labs or groups)	38%	30%	30%
Data citation services (e.g., assignment of permanent digital object identifiers (DOIs))	24%	42%	34%

Some respondents have previously submitted data to a repository, including the NCBI archives, especially GenBank; the Protein Data Bank; Cambridge Crystallographic Database; ICPSR; arxiv; as well as to several journals. A greater proportion of respondents in the sciences (45%) have contributed data to a repository than any other field. Overall, most respondents (60%) have not previously deposited their data into a repository, most commonly because they were unaware of any appropriate place to put the data (45%). A somewhat smaller group (26%) knew of places to put their data, but just had not had the time to do so.

What is the primary reason you haven't deposited research data into an archive or repository?

Answer Options	*Number of Respondents*	*Percent of Total Respondents*
Not aware of any appropriate place to put it.	14	45.2%
Just not had the time to get around to it.	8	25.8%
Do not want to share it publicly.	4	12.9%
Not produced research data worth preserving before.	3	9.7%
The conditions of use of my data do not allow me to deposit the data elsewhere or share it.	2	6.5%
My current research project is the first in which I'm generating / using data.	0	0.0%

Several chemists noted that though they do not put their data into repositories, they typically make their data available through publications, especially through supplements.

A majority of respondents (65%) said that their fields' major funders expect information about how the researchers will manage, preserve and/or share their data. Only 13% indicated that their funders do not expect this information. In the sciences and social sciences, a majority of respondents (83% and 67%, respectively) indicated that their funders require data management or sharing plans, compared to just over half (55%) in engineering. In the humanities, the largest proportion of respondents (33%) indicated that they were uncertain whether their funders required data management information or not.

Do major research funders in your field expect you to provide information in your funding proposals about how you will manage, preserve and/or share the data that you create during the course of your research? Please answer this question regardless of whether you are currently working on externally funded research or not.

Answer Options	*Number of Respondents*	*Percent of Total Respondents*
Yes, they do expect this information.	34	65%
No, they do not expect this information.	7	13%
There is more than one significant funding body in my field, and their requirements vary.	6	12%
I'm not sure.	5	10%

Most respondents (76%) were currently working on externally funded research, and of those, 72% had been required to include information about their data management practices in their grant. However, only a few respondents who had provided data management information had actually received feedback on their

plans (18%). About 32% felt that the data management plan had probably been considered, while 39% felt that it probably had not.

Do you think your funder seriously considered your responses before reaching a decision regarding the funding?

Answer Options	*Number of Respondents*	*Percent of Total Respondents*
Probably not, although I do not have direct evidence for that.	11	39%
Probably yes, although I do not have direct evidence for that.	9	32%
Yes, it was referred to in the feedback.	5	18%
No, I know for a fact that statements regarding data management did not inform the funding decision.	2	7%
Don't know	1	4%

Respondents indicated that in general, they were interested in sharing their data in the future, with or without restrictions. About 16% said that all of their data could be shared without restrictions; an equal number said that none of their data could be shared without any restrictions. A small number cited ethical or privacy concerns or other reasons that would prevent them from sharing any of their data; many more felt that some portion of their data could be shared, while other data would need to be restricted or otherwise held back.

In the humanities and engineering, no respondents reported that all their data could be shared without restrictions, while 21% and 22% in the sciences and social sciences said they could share all their data without restriction. More respondents in the humanities (33%), engineering (45%) and social sciences (44%) said there were ethical or privacy concerns that would prevent them from sharing some or all of their data, compared to less than 10% of respondents in the sciences. About 45% of engineering respondents also noted that some of their data could not be shared due to commercial or legal restrictions.

Would you be interested in the future in sharing your data?

Answer Options	*True for all of my data*	*True for some of my data*	*True for none of my data*
Yes, without restrictions	16%	65%	16%
Yes, but only in response to a written request explaining how the data would be used	25%	53%	20%
Yes, but only with colleagues or collaborators	24%	54%	19%
No, there are ethical and/or privacy concerns that preclude sharing	9%	38%	50%
No, there are legal or commercial restrictions preclude sharing	0%	59%	38%
No, there are other reasons I would not be interested in sharing my data	9%	26%	62%

About 34% of respondents have been inspired to undertake new research as a result of looking at shared data. These respondents primarily (41%) found this shared data through publications, either articles or books, rather than by browsing online or looking at data repositories.

How did you find out about the data that inspired you?

Answer Options	*Number of respondents*	*Percent of total respondents*
From a discussion or analysis of the data in a published article/book	16	41%
Found in a data repository	6	15%
Found on the Web, but not held in a data repository	6	15%
From a conversation with a friend or colleague	3	8%

Conclusions

This survey demonstrates that data management practices and needs at the University are diverse, but that dealing with data is a concern common to all fields. The respondents to the survey are a representative sample of faculty in Arts, Sciences and Engineering: humanists responded no less readily to the survey than scientists and engineers. This is an important point, given that similar studies conducted at other universities found that humanists were less likely to respond to queries about data management, and that they tended to be underrepresented in discussions of data services. It is clear that at the University of Rochester, humanists are engaged in data-driven research, and their needs should be considered when planning data management support and services.

Researchers expressed an open attitude toward sharing their data beyond their own collaborators. In some cases, ethical and privacy concerns limit the amount of data that can be shared or how widely it can be shared, but much of the data researchers currently manage could be shared without restriction—if researchers had easy-to-use channels for disseminating their data. Many have not shared data through a repository before, in part because of difficulty identifying an appropriate place to place their data, or because of the time and effort required to deposit data.

While researchers work with a wide range of data types, they express a shared need: somewhere to deposit data that needs to be preserved. Though not all data needs to be preserved indefinitely, current storage services are not meeting researchers' needs for the data they do want to save. The idea of an institutional data repository was strongly supported by faculty in all disciplines. Aside from infrastructural concerns, researchers expressed interest in education and training in data management, especially support in creating data management plans for funders. This is an area in which the Libraries are already actively working, and we will continue our efforts to increase services and build awareness of available support.

Overall, researchers at the University are actively engaged with data issues. The strong response rate suggests that researchers here are motivated to share their experiences with data management and help shape the future direction of data services on campus. Many are aware of their funders' requirements for data sharing and data management, and are looking for support in complying with those regulations, which makes this an opportune time to develop policy and services to suit their needs.