

Concordia University Libraries

Report on the 2016 survey of research data management at Concordia

Alex Guindon, Danielle Dennie
5/18/2017

Contents

Key findings.....	2
Survey	2
Interviews.....	2
Introduction	4
The Online Survey	4
Research team.....	4
Online survey	4
Methodology	4
Structure of the survey	5
Survey Results	5
RESEARCHERS' AFFILIATION.....	5
PART 1: DATA COLLECTION AND CURATION.....	6
PART 2: DATA MANAGEMENT.....	9
The Interviews.....	19
Research team.....	19
Interviews.....	19
Methodology	19
Interview results.....	20
PART 1 – DATA DOCUMENTATION	20
PART 2 – DATA STORAGE	20
PART 3 – DATA SHARING.....	21
PART 4 – DATA ARCHIVING	23
PART 5 – DATA WORKSHOPS	23
Conclusion.....	26
Appendix 1: Survey Questionnaire.....	27
Appendix 2: Research Data Management Interview protocols	39
Social Sciences, Humanities, Business, Fine Arts.....	39
Science & Engineering	43

Key findings

Survey

- Eighty-five percent (85%) of researchers have received funding from one of the tri-agencies and thus will be affected by their RDM policy.
- Since 66% of respondents have textual data and 68% numerical, Concordia or the Library should be able to provide tools/services based on existing RDM solutions. However, the University or Library may need to explore other solutions for the 47% of researchers who have multimedia data.
- Most researchers have documentation but the type of documentation varies widely from one researcher to the next. This speaks to a need for training to ensure adequate documentation. A Concordia or BCI data repository will need to accommodate the variety of documentation and metadata formats.
- Forty-five percent (45%) of researchers have large datasets that may be harder to deal with outside of disciplinary repositories that can accommodate very large files.
- Ninety-one percent (91%) of researchers store their data on PC hard drives. There is an urgent need to secure their data and allow them to easily share it.
- Systematic backup of data is only done by less than 50% of researchers.
- Over 70% of researchers want to deposit their data in a repository or learn more about how to do this.
- Forty-one percent (41%) of researchers want to keep their data forever, but retention policies should be put in place if institutional space was created for their data.
- The majority of researchers share their data, but few do so from a repository.
- Approximately a third of researchers submit articles to journals that require data, but the same proportion avoid these types of journals.
- Over 75% of researchers are interested in workshops for data management activities required by funding agencies.

Interviews

Data documentation

- Although in the survey, the majority of researchers said they had documentation for their data, many interview participants indicated that they either do not document their data, or that they consider the methodology, described in their published results, as sufficient to describe their data. Researchers in the humanities indicated that there was no incentive or reward for documenting or managing data. The two scientists interviewed were the only ones to document their data.
- In the 2013 survey, we interviewed social scientists. Most of them mentioned keeping some documentation, but this was not done systematically and the documentation was not always readily available or stored with the dataset.

Data storage

- A majority of researchers were struggling to find adequate space to store their current and past research data. Many researchers expressed a strong desire for a central university-run server, with clear policies and guidelines that could provide storage space for research data and

software to manage the data. This desire was expressed particularly by engineers and scientists. Researchers that had confidential data from human subjects, or secondary data providers, stated the importance of a secure space. Although 41% of survey respondents used cloud servers, some interviewees expressed security concerns over commercial servers based in the United States. A small minority did not want to store their data on university servers, or any servers in order to keep complete control over their data and any technical issues that may arise with a server.

Data sharing

- The majority of researchers interviewed, from all disciplines, were not sharing their research data, except with research collaborators or students involved in the project despite, in some cases, having received funding from agencies that have policies on open data. The survey showed that the majority of researchers shared their data, but the survey did not allow us to know with whom. Confidentiality requirements spelled out in ethics consent forms, was the most frequent reason mentioned as to why research data was not being shared. The perceived lack of research community need for raw data was also cited as the reason for not sharing. However, the vast majority of interviewees were willing to share their data because of its perceived benefits, although most would only share under certain conditions such as anonymization, controlled access, embargo periods, or the use of Digital Rights Management (DRM). Sharing secondary data was raised as a complex issue to solve. Only three researchers were currently sharing some or all of their research data because of perceived benefits or funder requirements.

Data archiving

- Of all the researchers interviewed, only the scientist was depositing data in an archival repository. Characteristics mentioned by most as being important in an archive included: 1) the ability to store data for very long time periods (especially helpful for hard to collect data, data useful for comparative studies in time, aggregating data from multiple research, or long-term data mining); 2) stable funding (preferably from funding agencies) to ensure the permanency of the archive; 3) policies surrounding data preservation and access; 4) and mechanisms to ensure that the data is secure. A minority of respondents did not see archiving as necessary, stating there was no research or community need to reuse data or to reanalyze it.

Introduction

This report presents the highlights a survey of research data management (RDM) practices and needs conducted at Concordia in the fall of 2015 and winter of 2016. This survey, and a series of follow-up interviews, was directed at the complete faculty population of Concordia University with the exception of three departments, Geography, Political Science and Sociology which were already consulted in a previous RDM survey conducted in 2013. The faculty from the Department of Psychology which were included in the 2013 survey, were surveyed again due to a particular interest in that group from Dr. Jennifer McGrath, one of the co-investigators.

The first section of the report deals with the results of the survey and the second section discusses the main findings of post-survey interviews conducted with researchers. In the conclusion, we discuss the main obstacles to adopting RDM best practices and upcoming developments that can inform the development of RDM services at Concordia University. We hope that this report, in conjunction with the *Tri-Agency Statement of Principles on Digital Data Management*¹ will demonstrate the need for the Libraries, and for Concordia as a whole, to tackle the pressing issue of RDM.

The Online Survey

Research team

The research team was composed of Alex Guindon, GIS and Data Services Librarian; Danielle Dennie, Reference Librarian; Jennifer McGrath, Associate Professor, Psychology. Muhammed Idris, a Post-Doctoral student working with Dr. McGrath and Dubravka Kapa, Associate University Librarian, Research and Graduate Studies were also consulted in the creation of the survey questionnaire and the general planning of the research project.

Although the survey instrument was a joint effort by all the people mentioned above, the qualitative part of the study –the series of in-person interviews and analysis thereof—was conducted exclusively by Danielle Dennie and Alex Guindon.

Online survey

Methodology

The objective of this research project was to obtain a picture of current Concordia faculty practices in terms of research data management (RDM). At the same time, questions were asked to assess faculty needs for assistance in the field of data management. The questionnaire (Appendix 1) was freely adapted from the Data Asset Framework (DAF) created by the Humanities Advanced Technology & Information Institute (HATII) at the University of Glasgow and supported by the Digital Curation Centre (<http://www.dcc.ac.uk/resources/tools/data-asset-framework>). Several questions were added to the questionnaire used in the 2013 survey to better assess the needs of faculty in the field of data science. These questions mostly appear in the WORKING WITH DATA: STATISTICS, PROGRAMMING, & VISUALIZATION section of the survey.

¹ <http://www.science.gc.ca/default.asp?lang=En&n=83F7624E-1>

The population of the Survey was constituted of all full-time, current faculty members except for the Departments of Geography, Political science and Psychology for reasons mentioned above. The online survey, hosted on Survey Monkey, was conducted from December 2015 to January 2016. The participation rate was 19%, with 132 participants out of 696 eligible faculty. However, most questions were not compulsory so the actual response rate by question varies.

Structure of the survey

The questionnaire was organized in a few different thematic sections. After some questions on researchers' affiliation (faculty, department, research team), the first section dealt with **data collection and curation**. Here, participants were asked about research funding, and about the type of data they gathered and who was responsible for managing their data. The second part of the questionnaire was called **data management**. Researchers were asked about all aspects of their data practices: the type and size of their dataset; where they stored it during the research project; where they would consider archiving it; and would they be interested in sharing their data, if so, with whom? The 3rd section -- **working with data: statistics, programming, & visualization**— was concerned with the tools and techniques used for data analysis and visualization. Finally, the 4th section --**data workshops and assistance**-- aimed at assessing the needs of researchers in terms of RDM assistance and the type of workshops they, or their research assistants, would benefit from.

This report only presents the highlights of the results; the complete results will be made available on Spectrum.

COMPARISON WITH THE 2013 SURVEY

For some survey questions, even though the exact text of the questions and the answer choices differed, we found it useful to make comparisons with the **2013 RDM survey**. Such comparisons appear in grey text boxes like this one.

Survey Results

IMPORTANT METHODOLOGICAL NOTE

Unless otherwise specified, the numbers reported are the percentage of respondents that provided a given answer. As most questions allowed for multiple answers, these percentages add up to **more than 100%**. In other words, the categories are not mutually exclusive.

RESEARCHERS' AFFILIATION

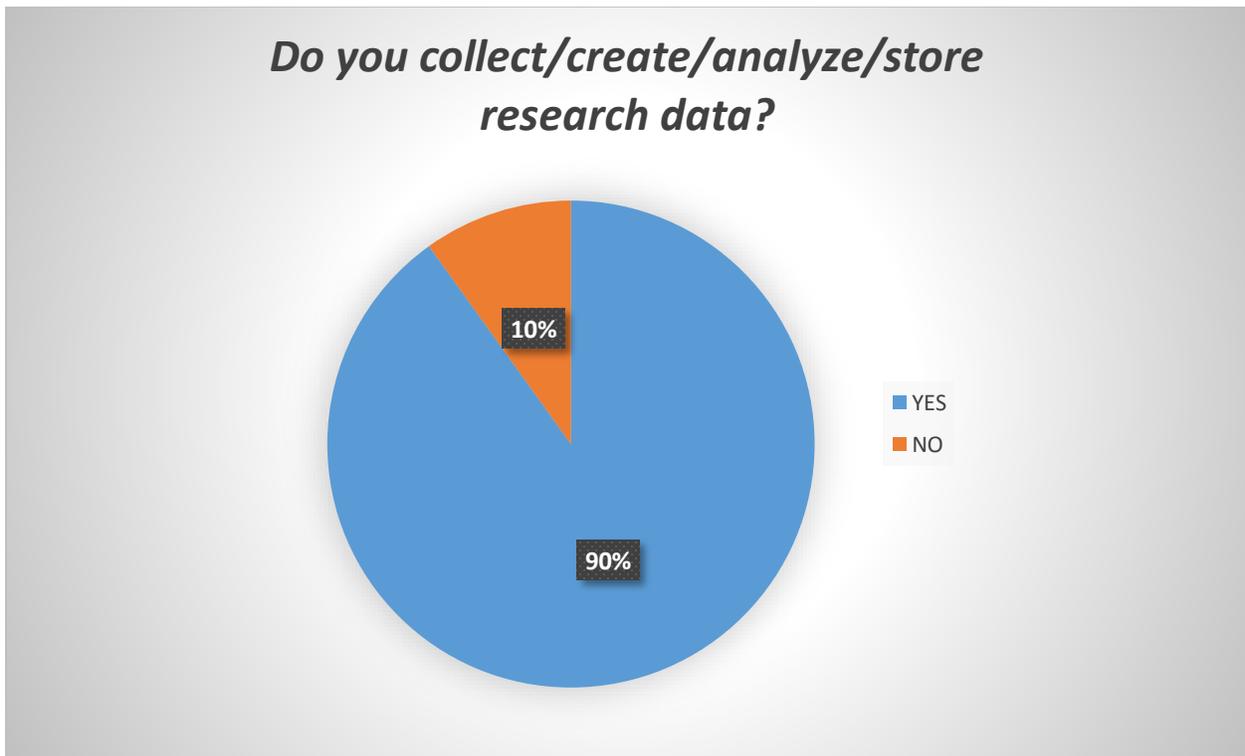
Most of Concordia's departments were represented in the survey but a large number of respondents were from either the Department of Psychology (17) or Exercise Science (11). This is most probably due to the presence on our team of Dr. McGrath whom those researchers know personally. This results in an overrepresentation of those two disciplines and it is possible that certain results are skewed towards the

answers given by faculty from those departments. Overall, we do not think however that the responses from these faculty members differed from other respondents in their views and practices of RDM.

In terms of affiliation with a lab or research team, we had a wide range of answers, with PERFORM (13) and the Centre for Clinical Research in Health (CCRH) (4) well represented.

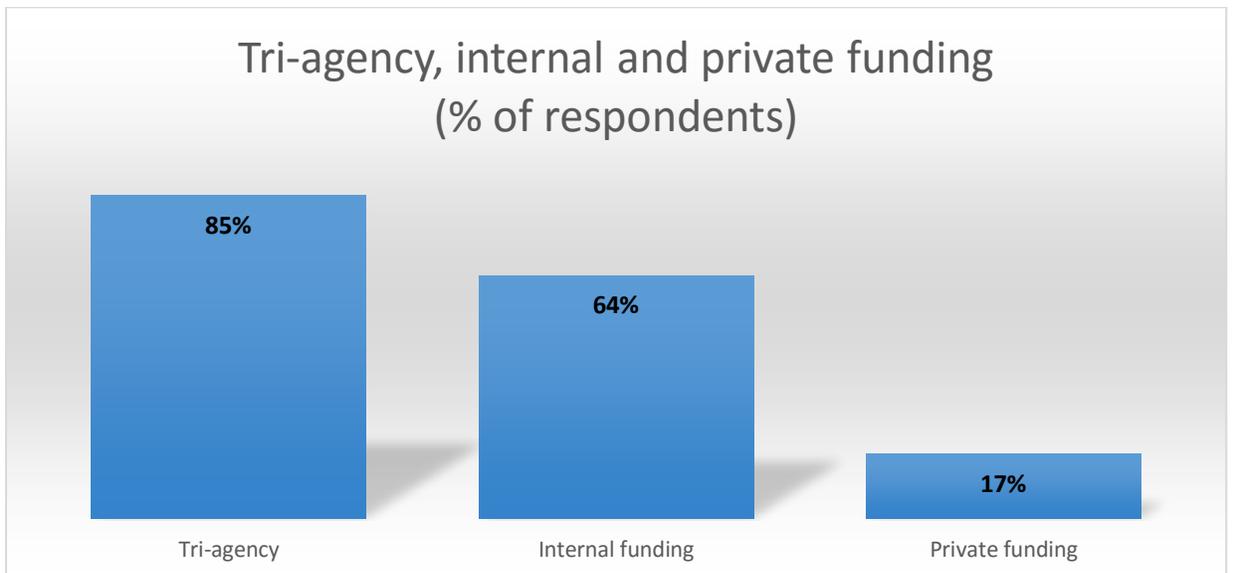
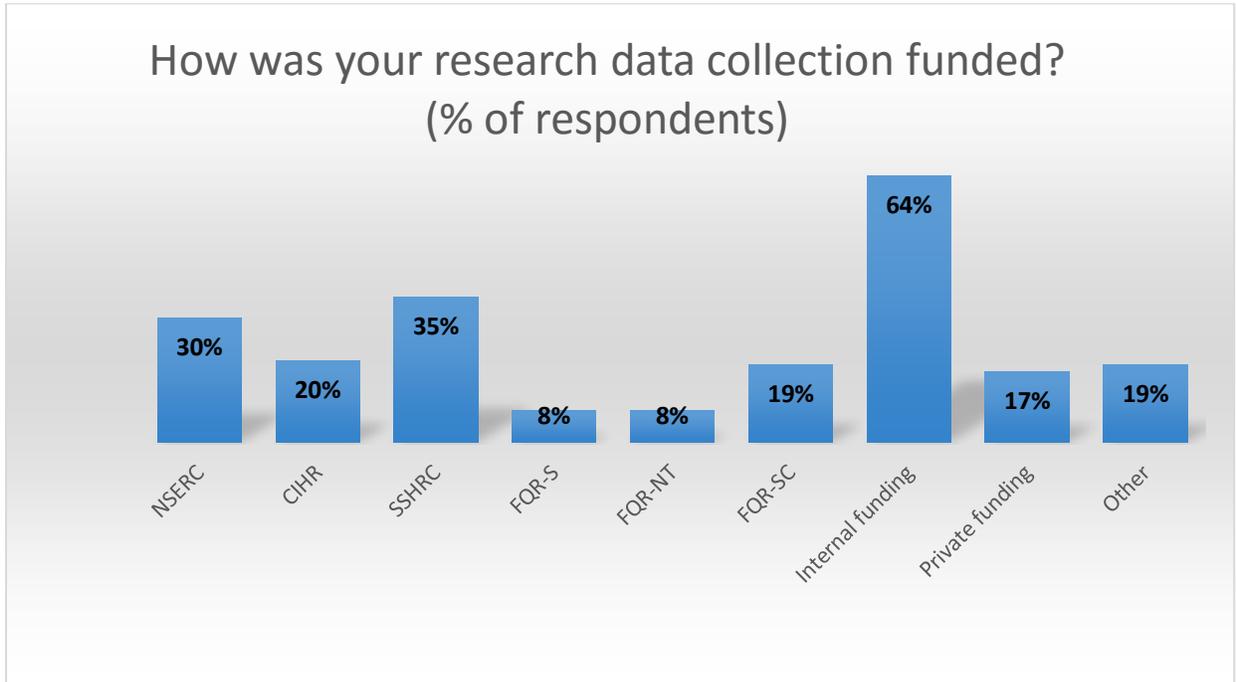
PART 1: DATA COLLECTION AND CURATION

Do you collect data?

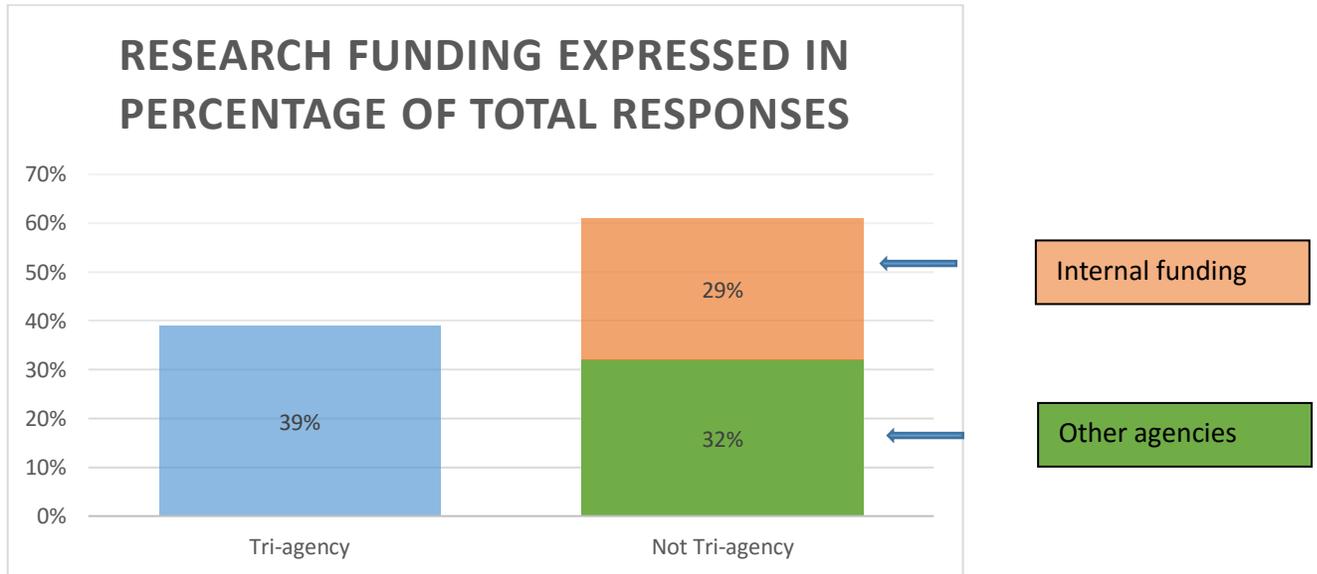


Data was defined broadly to include “all primary data collected during your various research projects at Concordia. This might include survey data, experimental data, simulation data, programming code, qualitative data (interview transcripts, field notes, audio or video files), or other information.” Not surprisingly, 118 respondents out of 131 answered that they indeed collected data. This shows that a large cross-section of Concordia researchers would benefit from RDM services.

Research funding



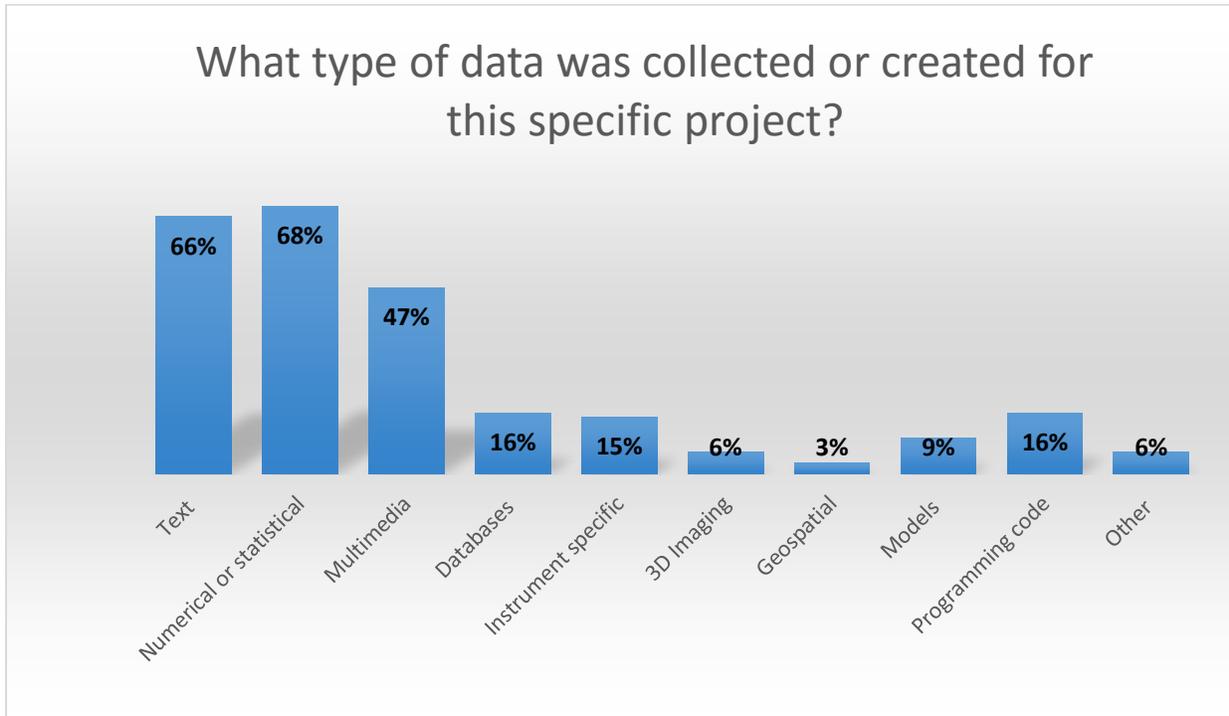
We could obtain a more complete and more accurate breakdown from the Office of Research, but judging from the people who answered our survey –and who are presumably interested in RDM—we can see that several researchers (15%) will not have to abide by the forthcoming RDM requirements of the tri-agency. On the other hand, these researchers are less likely to be able to request grant money for the purpose of data management.



Since respondents were allowed to choose multiple answers, another way to analyze the results is to count the total number of responses. The number of responses is a better approximation of the distribution of grants as the same researcher can have received several grants for various projects or even for the same project. It is interesting to note that 61% of the grants come from funders other than the three large federal ones (tri-agency). Indeed, internal funding is important at Concordia with 29% of grants reflected in this survey. Particular attention will be needed to support datasets created with the support of private funding (8%) as these are often by special license agreements with the funders and may, for instance, require complete confidentiality.

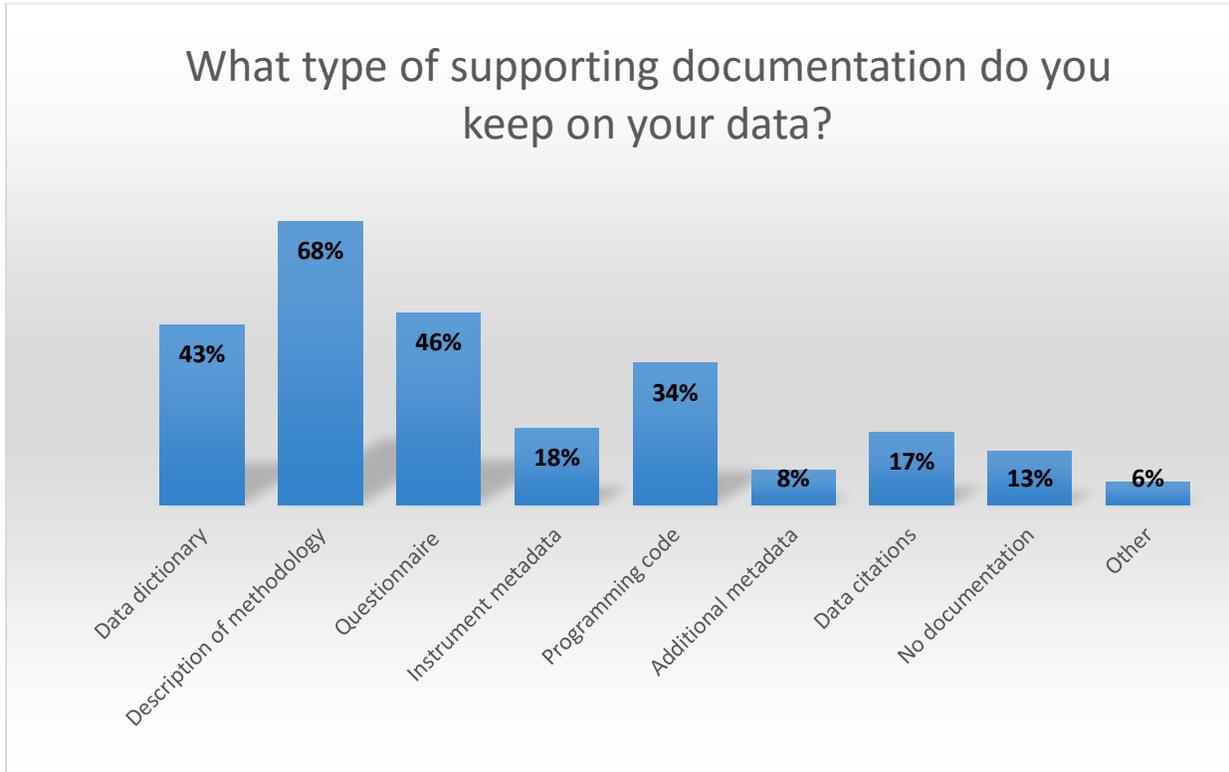
PART 2: DATA MANAGEMENT

What type of data was collected or created for this specific project?



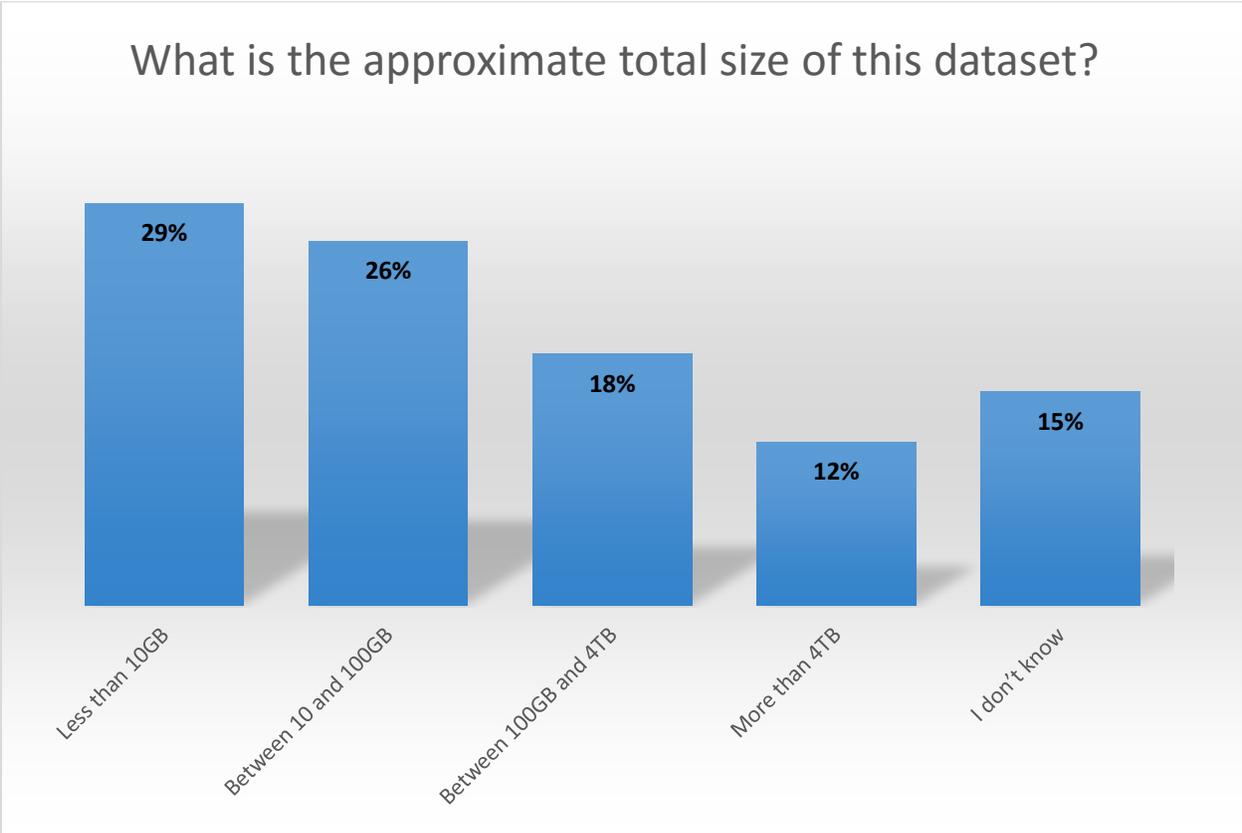
The results show a very wide range of data type, the most common being numerical (68%), text (66%) and multimedia (images, audio, video) at 47%. Data services and tools should be, as much as possible, designed to accommodate this variety of data types. More advanced services and tool functionalities should probably be geared at statistical and textual data. Multimedia files are very different than the two previous categories and may necessitate different tools and services.

What type of supporting documentation do you keep on your data?



The results show that researchers keep a large spectrum of documentation on their dataset. Only 13% do not keep documentation. The type of documentation most likely varies depending on the type of data described. For instance, only experimental data would require instrument metadata. This variety of metadata implies that we choose flexible tools for a data repository. But as mentioned regarding the type of data being collected, it is likely that certain types of documentation would be better accommodated by disciplinary platforms when they exist.

What is the approximate total size of this dataset?

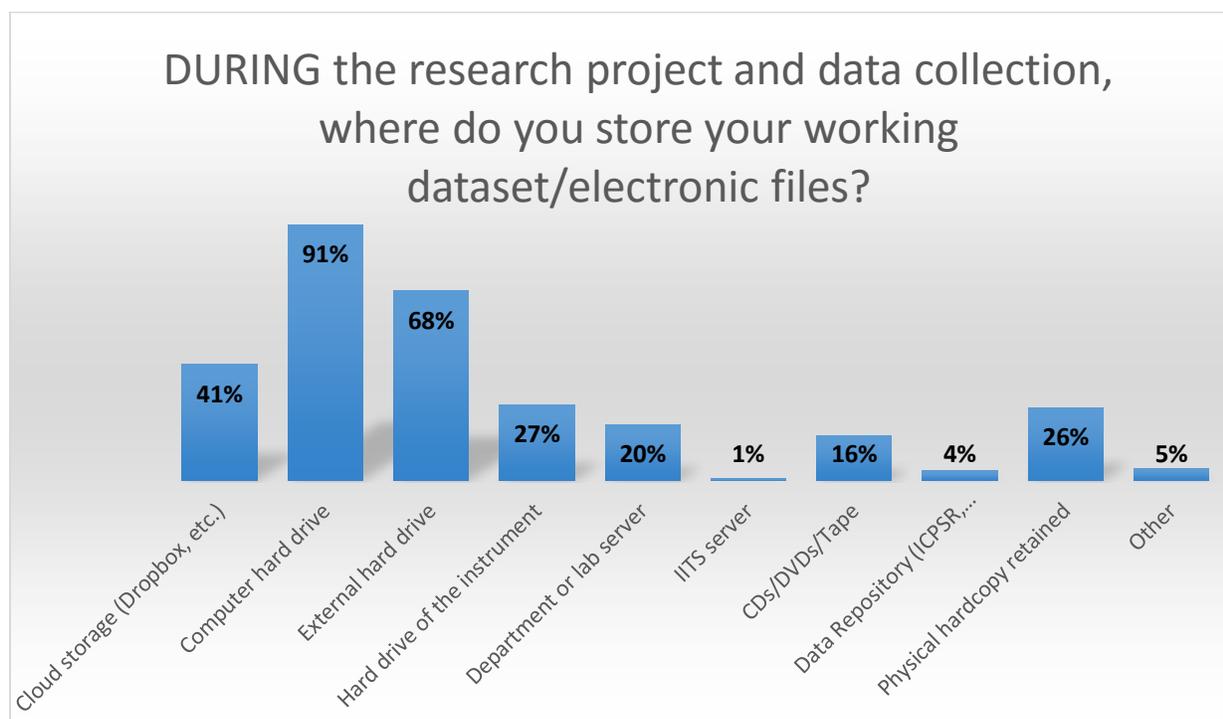


Fifty-five percent (55%) of the respondents reported that their most important dataset was less than 100 GB in size. Most data repository platforms (like Dataverse for instance) should be able to accommodate those datasets. That still leaves 30% of researchers with large datasets (plus 15% reporting datasets of unknown size) that may be harder to deal with outside of disciplinary repositories built to receive massive data. More analysis would be needed to determine if those large datasets come mostly from the natural sciences.

COMPARISON WITH THE 2013 SURVEY

In 2013, only 11% of respondents claimed to have datasets larger than 100 GB (although as many as 30% did not know the size of their dataset). This seems to indicate that the vast majority of social science datasets can be handled by the common data repository platforms.

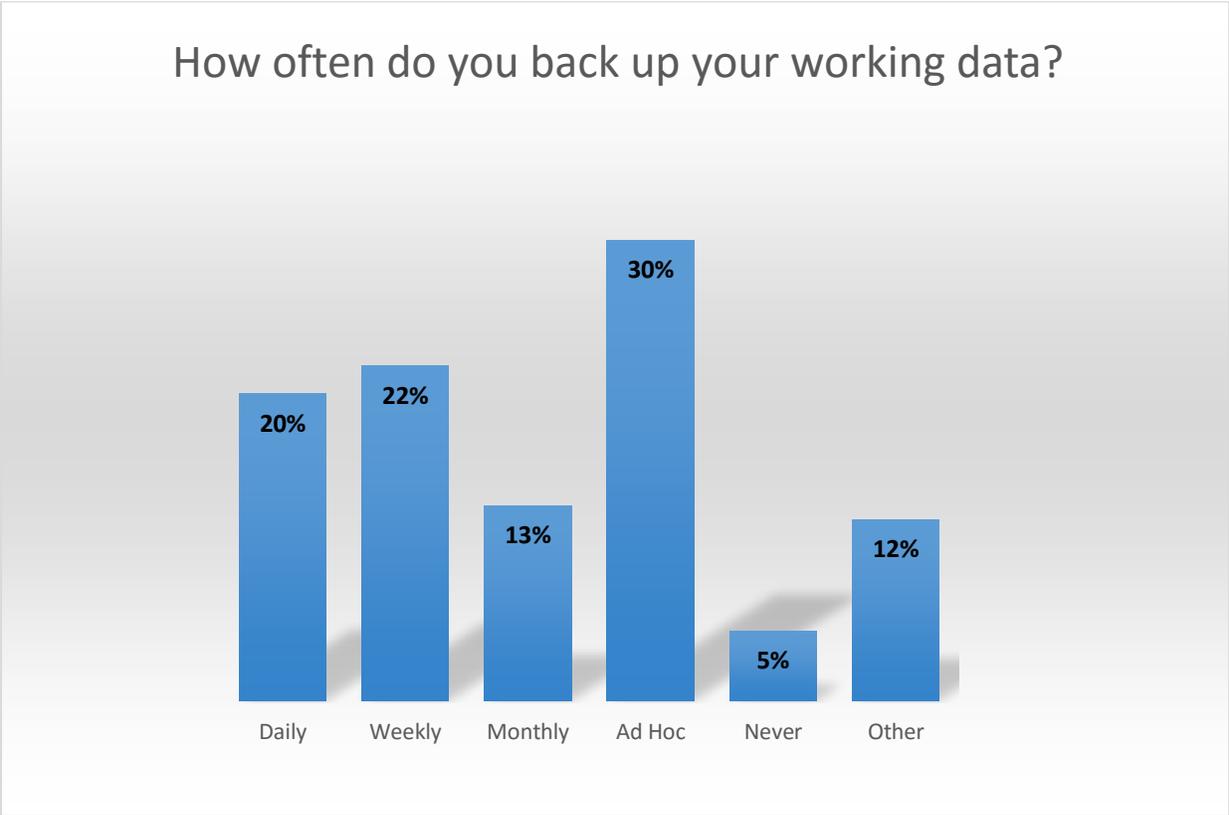
DURING the research project and data collection, where do you store your working dataset/electronic files?



The answers to this question show a wide range of options chosen to store the active dataset(s). As 91% of respondents use their PC's hard drive and only 20% mention a department or lab server—1% indicate an IITS server—we can assume that most researchers do not benefit from an ideal way to secure their data and share it with other team members. The use of cloud storage (41%) is common and may facilitate data sharing with co-researchers but may pose security and confidentiality risks. These results show a need for a Concordia-based network storage or *virtual research environment* that would focus on data security and secure sharing capacity.²

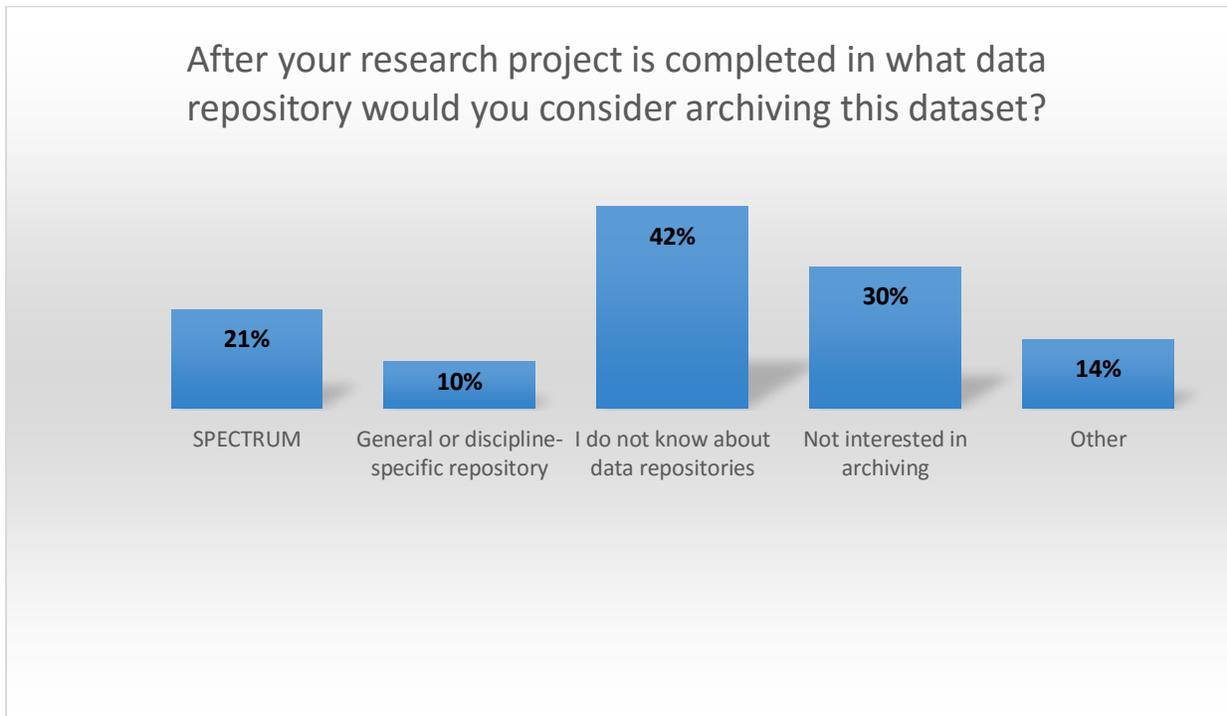
² IITS has conducted a survey of large research centres at Concordia and obtained similar results. They plan to offer some technological infrastructure to those centres that could then be extended to individual researchers outside of those units.

How often do you back up your working data?



Only 42% of respondents backup their data at least once a week. This shows a significant potential for data loss. The fact that 30% of researchers indicated using an ad hoc method back up strategy demonstrates the lack of formal data security protocol. Some training and basic tools would contribute to ensure a much more systematic back up procedure.

After your research project is completed –the data has been cleaned and analyzed— in what data repository would you consider archiving this dataset?

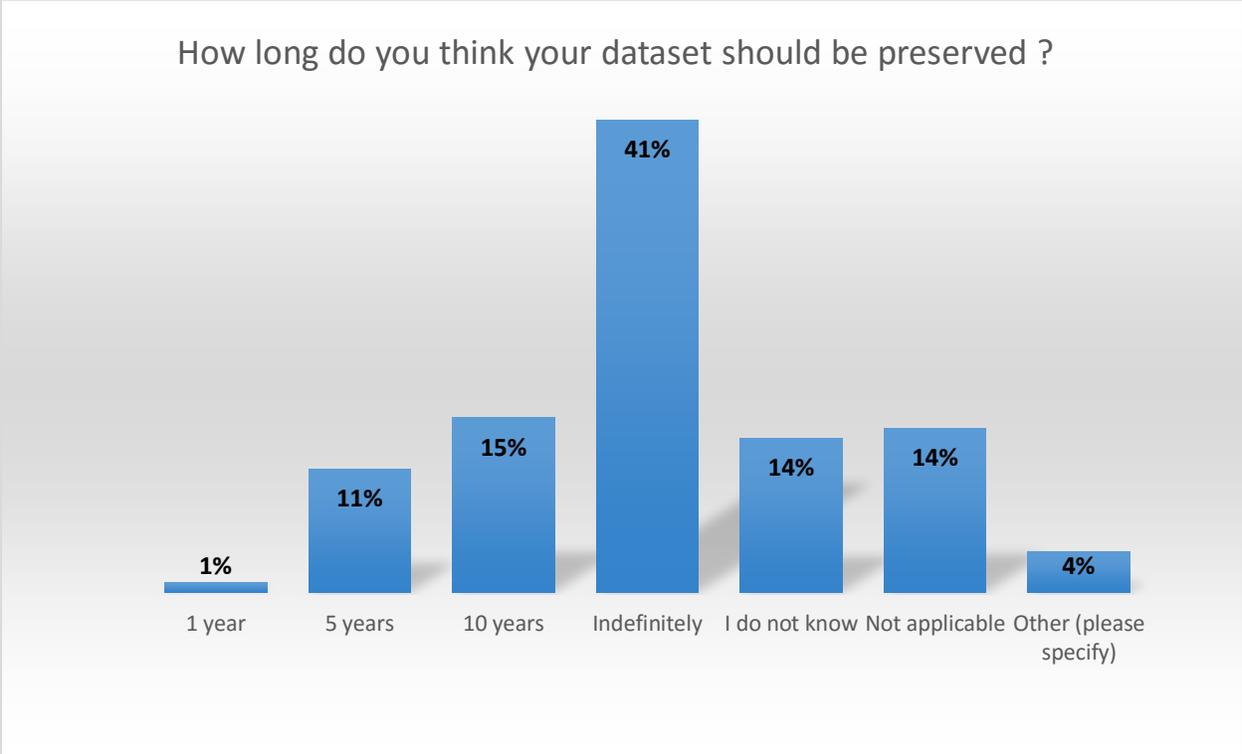


Although there is nearly a third (30%) of researchers who indicate that they are not interested in archiving their data –there may be a wide range of reasons for this attitude, ranging from a restrictive interpretation of ethics compliance to a lack of understanding of the advantages of data archiving (and data sharing)—a significant share (31%) would rather archive it (in Spectrum or another data repository) and an even larger number (42%) are at least keen to learn more about archival options. It is remarkable that twice as many respondents showed a preference for a local archiving option (Spectrum) over a general or discipline data repository. This likely reflects either a confidence in Concordia’s (or Concordia Library’s) ability to provide a safe place for data archiving or the researchers’ attachment to their home institution, or possibly both. On the other hand, there is probably a lack of awareness of some of the large, state of the art repositories such as ICPSR in the social sciences.

COMPARISON WITH THE 2013 SURVEY

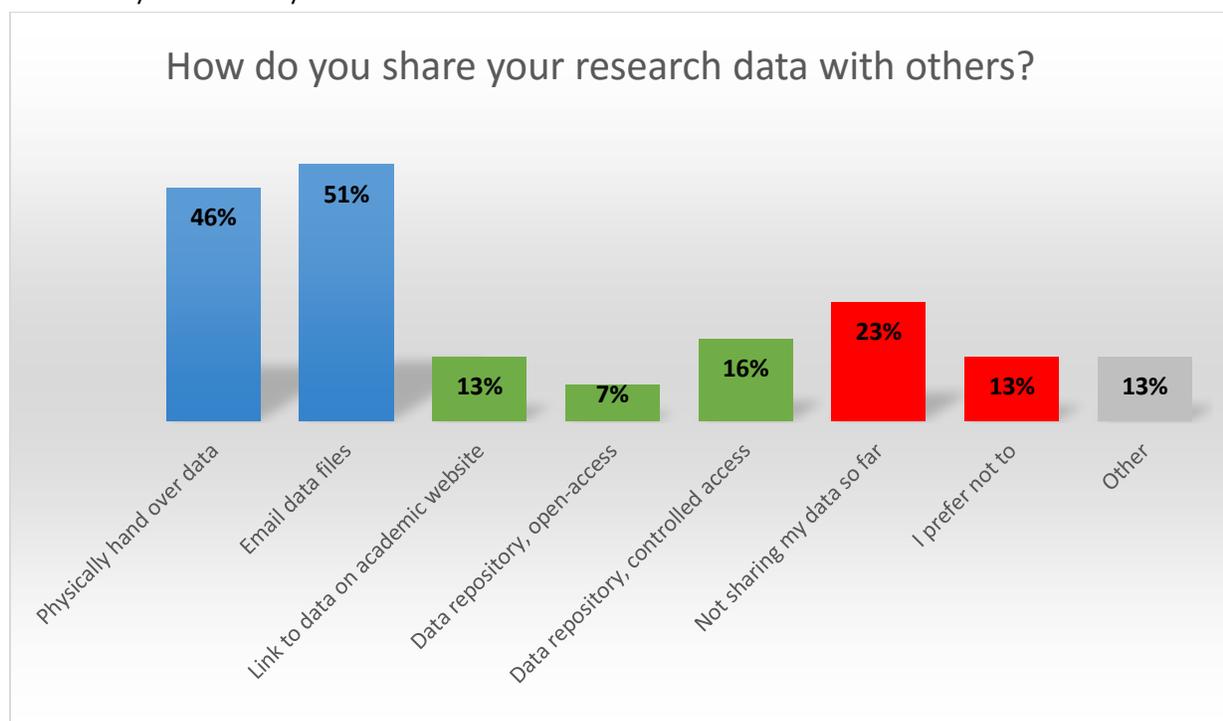
These results are consistent with those of the 2013 survey. In that survey, 32% of respondents said they were not interested in long term archiving of their data. The choice of answers regarding data repositories in the 2013 survey was different, but a significant number of people (39%) expressed interest in Spectrum, while 29% said they were interested in an external data repository.

How long do you think your dataset should be preserved in a data repository after the completion of your project?



There is a strong preference for archiving the data indefinitely. Although this may be desirable for many datasets, it may not be necessary –or even technically feasible-- in all cases. Clear data retention policies should be devised in accordance with the granting agencies’ policies. As reflected in the answer to this question, some datasets may lose their usefulness after a certain period.

How do you share your research data with others?



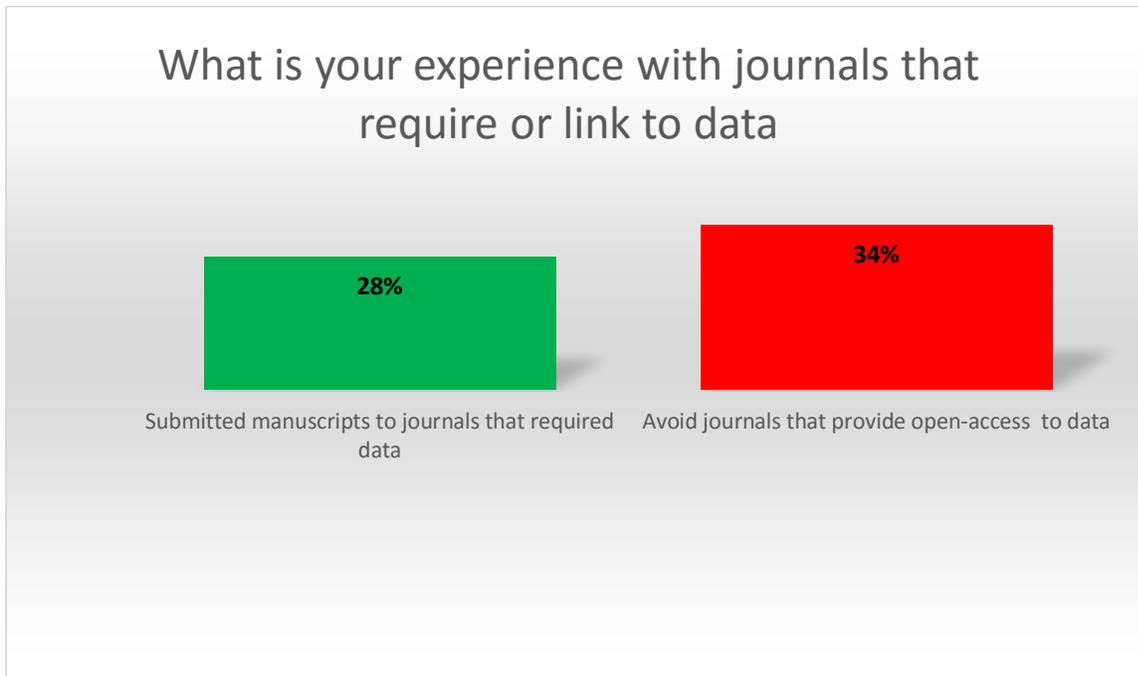
- Shares data manually
- Shares data through data repository or website
- Currently does not share data

Only 13% of researchers answered that they prefer not to share their data. Twenty-three percent (23%) indicate that they have not shared their datasets so far, but would like to learn more about data-sharing options. This is very encouraging from the perspective of open science. However, only a minority of researchers have used data repositories (7% open access, 16% controlled access) for sharing their assets. To this group, we may add 13% that link to datasets on an academic website (which may or may not be considered a data repository). The rest rely on a mix of ad hoc methods like physically handing over data or emailing them which are obviously not ideal in terms of data security or from a practical standpoint. This way of sharing data probably reflects the researchers' will to maintain control over data dissemination.

COMPARISON WITH THE 2013 SURVEY

In 2013, 31% of researchers were ready to make their dataset publicly available, 24% preferred to use a controlled-access platform and 30% mentioned that they would personally grant access to their data. Finally, 15% of respondents said that they were not interested in sharing their data. Given the different choice of answers, the only meaningful comparison is the percentage of people who are not interested in sharing their data (15% in 2013, 13% in 2016). We can reasonably conclude that a large majority of researchers are open to sharing their data although the preferred conditions through which this dissemination would take place vary widely.

Thinking about peer-review journals in your discipline, or other respected sources for research dissemination, what is your experience with the following?



There is a growing number of journals that require researchers to provide the dataset associated with their paper. This is reflected in the relatively high number (28%) of respondents who have indeed submitted manuscripts to such publications. However, there is an even larger number (34%) of faculty who have avoided publishing in journals that have an open-access data policy. Note that not all journals that require data files necessarily make them available in an open-access way. For instance, some may allow for controlled access to data files.

For each of the following, please indicate your level of interest in data management assistance at Concordia or training/development opportunities

Workshop/Resource	Very interested	Somewhat interested	Total interested	Not interested
Data management activities that will be required by grant funding agencies	43%	34%	77%	15%
Data management plan	30%	42%	72%	21%
Research collaboration software tools	29%	39%	69%	25%
Finding and accessing existing data sources	29%	34%	63%	23%
Data archiving	27%	43%	69%	22%
Data visualization	27%	44%	71%	22%
Digitization of paper records	27%	39%	65%	26%
Data sharing and access	23%	40%	63%	30%
Help in ensuring the confidentiality/anonymity of data on human participants	21%	28%	49%	29%
Data management workshop	19%	49%	68%	25%
Data documentation and metadata	18%	38%	56%	33%

These results are presented in descending order for the “very interested” category. By far, the most desired category for assistance is “activities that will be required by grant funding agencies”. While not a surprising result, it reinforces the generally accepted perception that RDM will become a much more important issue in academia once clear requirements from granting agencies come into force. In general, it appears that RDM assistance will be desired for a wide spectrum of activities. Even the least popular topic in the “total interested” column (help in ensuring the confidentiality/anonymity of data on human participants) shows a score of 49%. Providing assistance on all these aspects of RDM will be challenging and will most likely require the participation of several stakeholders.

The Interviews

Research team

The research team was composed of Alex Guindon, GIS and Data Services Librarian and Danielle Dennie, Reference Librarian. Graduate students were hired to transcribe the interviews.

Interviews

Methodology

While the online survey provided an overview of current research data management practices at Concordia and some basic information on important datasets held by researchers, it was not sufficient to obtain a detailed description of those datasets and to discuss the specific needs of faculty for data archiving and sharing. This type of in depth information can only be gathered through in-person interviews and discussions with researchers.

Ten interviews were conducted with faculty who indicated –in answer to one of the survey questions– that they would accept to be contacted for further discussion. Researchers were chosen based on their level of interest in archiving and potentially sharing their data. Researchers from the following department were interviewed:

- Building, Civil and Environmental Engineering
- Design and computation arts
- Computer science
- Physics
- Exercise science
- Management
- Supply Chain & Business Technology Management
- Communication studies
- Studio Arts
- Études françaises

The interviews were one hour long, semi-structured and, like the online survey, loosely based on the Data Asset Framework from the Humanities Advanced Technology & Information Institute (HATII) at the University of Glasgow. The two interview protocols (one for the science disciplines and the other one for social sciences, business and fine arts) can be found in Appendix 2. Analysis of the interview transcripts was done using a two-step process. During first cycle coding, descriptive coding was used to code the transcripts³. During second cycle coding, focused coding was used to analyze the data⁴.

³ Johnny Saldana, *Coding Manual for qualitative researchers* (Los Angeles: Sage, 2009), 70-73.

⁴ *Ibid.*, 155-159.

Interview results

PART 1 – DATA DOCUMENTATION

From the results, it seems that inexistent metadata is the norm across all disciplines. The missing metadata is explained by the fact that some believe that their **data is self-explanatory** or sufficiently clear to someone in the field, while others consider the **methodology**, described in their published results, as **sufficient to describe their data**. A few researchers also mentioned the fact that there was **not enough time** to adequately document their data. In fact, the engineer noted that “documentation is gonna be (...) extra work for me and my team. And so, (...) I’m not really sure if you want to go that way because I prefer students to do scientific work and publish papers then, you know, be data managers.” Finally, the two humanities researchers noted that there are **no incentives or rewards** for properly documenting (or managing) data as there are for other parts of the research process, such as publishing. On the topic of documenting datasets, one humanities researcher noted “If there was acknowledgement of sustainability and there were rewards for that, that would be fantastic. There’s more rewards for bringing new stuff in then there is rewards for careful preservation. And that’s my observation from 25 years here.”

Researchers who use **secondary data** (engineering and business) noted that they don’t need to create metadata as documentation was already created by the original owner of the data.

The **only researchers to document their data are two scientists**. The physicist keeps read-me files and developed a file naming convention that describes the data. The computer scientist, working in bioinformatics, deposits research data in open repositories which use metadata to describe data files. Furthermore, to manage the large amount of samples and files generated by the instruments, the computer scientist uses a laboratory information management system (LIMS) which keeps metadata on samples and from the instruments.

COMPARISON WITH THE 2013 SURVEY

In the 2013 interviews, most social scientists reported keeping some documentation although it was never done in a systematic way. Some respondents mentioned that the documentation would have to be located and somehow associated with the dataset.

PART 2 – DATA STORAGE

The interviews made it clear that a majority of researchers were struggling with finding adequate space to store their current and past research data. Many researchers expressed a **strong desire for a central university run server** that could provide storage space for research data. This desire was expressed particularly by engineers and scientists. Researchers that had **confidential data** from human subjects, or secondary data providers, stated the **importance of a secure space**. A few researchers also mentioned the **importance of getting technical support** if server space was provided. One of the scientists summed this up quite succinctly during the interview:

We need an institution server (...). [W]e should have secured access with secured account for each research team, for each lab. (...) [T]here should be a flexibility [to accommodate] people with space needs. [We] need (...) a technical team supporting this server (...) someone to (...) check that there is enough available space, see if there's problems, to fix the problem, to manage the access, the accounts, the codes, and make sure [it is] regularly backed up and whether there should be some (...) duplicate somewhere, to make sure everything's safe.

Alongside the need for university run server space, a few researchers indicated that they would benefit from **university created policies and guidelines on how and where to store research data**. Some researchers were notably nervous about commercial cloud servers based in the United States. The humanities researcher noted: "I think we need our own indigenous resources, so that (...) we're under the injunctions of the agencies that are federally funded and mandated, yet the problem of data leakage or data surveillance or data theft, once it is outside of our borders, it's in some ways beyond our control (...)."

Researchers in the fine arts and humanities wished for **software to help better manage stored data**. The software would allow **organizing, tagging and visualizing stored data**.

Finally, **not all interviewees requested university servers for storing research data**. Concerns were raised about the ability of IITS to manage research data servers or to provide adequate and timely technical help. One researcher in particular wanted complete control over data produced during the course of the research and felt that having a server owned by the laboratory would best suit this need.

Ça c'est une des raisons pour lesquelles le labo avait ses propres serveurs, c'est que on s'inquiétait un peu des délais pour pouvoir régler les problèmes et (...) des délais pour régler les problèmes (...) ça entrave les recherches, ça retarde les recherches et tout, donc, on veut le contrôle, (...) sur nos archives.

Another researcher preferred not have any data on a server, for security reasons. "I have concerns that keeping [my data] not on a server at this point, but keeping it on a hard drive, (...) makes me feel like I'm not going to have that information accessed by, you know, corporate media or a government agency that hasn't been given the ok to have access."

PART 3 – DATA SHARING

The **majority of researchers** interviewed, from all disciplines, **were not sharing their research data**, except with research collaborators or students involved in the project. In two instances, data was not being deposited in repositories for sharing **despite having received funding from agencies that have policies on open data**. In fact, one science researcher stated that they were able to "get out of these rules, because of the confidentiality, because of the ethical regulations."

Indeed, for researchers in all disciplines, except engineering, **confidentiality requirements** spelled out in ethics consent forms, was the most frequent **reason mentioned as to why research data was not being shared**. For one researcher in the John Molson School of Business, **funder policies on open data could possibly impede research by causing fewer people to want to be interviewed**. This researcher believed that the effect of a funder mandated open data policy would need to be measured.

Confidentiality requirements in ethics forms were not the only reasons for not sharing data. For two researchers, in science and fine arts, the **perceived lack of research community need for raw data, was cited as the reason for not sharing.**

Instead of sharing data, the best way to share research was through open access journal articles, according to one of the science researchers. “I’m actually happy that Concordia does cover (...) up to a certain extent, the cost of this open access [publishing]. When I can, I do publish in open access because I think that’s the best way to make knowledge accessible”.

Despite the vast majority of researchers not currently sharing their data, **all except one**, from the School of Business, **were willing to share their data.** In fact, a few researchers mentioned the **benefits** of sharing, such as the **ability to verify the analysis or reproducibility of the research** and **receiving citations to the data.** One researcher from the School of Business stated: “get[ting] citations (...) is (...) how my university recognizes the value of what I’ve done. (...) I realize that if your article is valuable, people might cite it, but a lot of the time, the most valuable thing about an article is the data that was collected.”

Despite the vast majority of interviewees saying that they were willing to share their research data, **many would only share under certain conditions.** These conditions were anonymization of or controlled access to the data, embargo periods on the data to allow the pursuit of patents, and finally the use of DRM to control how the data is used. For one of the fine arts researchers, “the big issue is making sure whatever is in the consent is honored.”

Finally, a few researchers remarked on the **complexities of sharing secondary data** (i.e. data compiled from secondary sources), which would involve getting the right permissions and ensuring that researchers using this data cite the correct data sources.

Of the ten interviewees, only three researchers, from computer science, humanities and fine arts, were **currently sharing some or all of their research data.** The reasons for sharing were because of **perceived public or research community interest.** In fact, in the case of the computer science researcher, sharing data was undertaken in order to follow “research community norms (...) which [are] very strong in promoting open data and open access”, as well as because of the **data release policies spelled out by the granting agency.**

COMPARISON WITH THE 2013 SURVEY

In the 2013 interviews, all researchers were willing to share their data. However, some researchers dealing with qualitative data in the form of interviews or observational videos were concerned with the difficulty of de-identifying their data. Importantly, none of the researchers had included the possibility of data-sharing in the consent forms that their research subject had to sign.

PART 4 – DATA ARCHIVING

Of all the researchers interviewed, **only the computer scientist was archiving** a portion of research data in public repositories, as mandated by the funding agency. Of the **nine other researchers**, although two were making their data publicly available (see previous section on sharing data), **none were depositing their data in archival repositories.**

Of the researchers interviewed, it seems as though **four features should be present for data archives:** the ability to store data for very long time periods, stable funding to ensure the permanency of the archive, policies surrounding data preservation and access, and mechanisms to ensure that the data is secure.

The researchers in the sciences, the humanities, and the School of Business all wanted **long term preservation of their data**, ranging from 10 years to indefinitely. Reasons cited for keeping the data were either because the original data was hard to collect, or because the data would be useful for comparative studies in time, for aggregating data from multiple research, or for continued data mining. In fact, one fine arts researcher stated: “I can probably keep five Ph.D. students busy for the next seven years just going through that old data, not even talking about the work we’re doing now.”

Two researchers who are currently openly sharing their data, in humanities and computer science, bemoaned the fact that **funding agencies do not provide funding or initiatives for data preservation.** “[T]he funding doesn’t work that way. [They] pay you to do the project, and once the project’s over, [y]ou [get] zero support. And if the data is still there when someone starts up the machine the next time, good, (...).”

The two researchers from the School of Business noted that **policies are needed regarding data archiving.** Policies could be written on the types of data deposited or how the data would be preserved, accessed or secured. On these topics, one researcher mentioned the complexities of archiving someone else’s primary data without getting explicit permissions. Another researcher was worried about confidential information and data breaches. From the previous section on data sharing, it is obvious that the researchers working with confidential data would like an **archive to provide controlled or very secure access.**

Finally, **only two researchers**, one in the sciences, the other in the fine arts, **were not interested in archiving** their data as they did **not perceive any research or community need.** Both researchers noted that most researchers in the field did not reuse data or need data for reanalysis.

PART 5 – DATA WORKSHOPS

Most interviewed participants indicated that many or most of the workshops would be of interest. In fact, one researcher suggested that it might be best to combine all of these into a course:

“[A]s a researcher that really cares about data, I just feel like all of this is absolutely key, but I don’t know when I’d have the time to do it. That’s why I think this needs to be a boot camp, or it needs to be a one week course. I think that if it was presented (...) in the kind of right way, even giving people a certificate at the end of it so that they can say they have (...) completed a course

in data management would be really good for somebody's CV even, making it so that [there is] official acknowledgement of certain kind of skill acquisition."

When asked to indicate specific workshops that would be of interest, the different interviewees chose the following:

Workshops	Interviewees interested	Possible reason for interest
Data management workshop (examples of topics: setting up electronic files, file formats, IP and licensing, de-identification of data)	Science (1) Fine arts (1)	Work with large audio, video, interview, instrument datasets
Data visualization (interactive graphics, etc.)	School of Business (1) Humanities (1)	
Data management plan (organization and curation of data throughout research project)	Science (1) Humanities (2)	
Data management activities that will be required by grant funding agencies	Science (1) Engineering (1)	Tend to get grants from big agencies (NSERC)
Data documentation and metadata (i.e. data dictionaries, data codebooks containing description of dataset and variables; other supporting documentation)	Humanities (1) Fine arts (1) Science (1)	Work with large audio, video, interview, instrument datasets, some with confidential information
Data archiving (e.g. finding appropriate data repository, understanding how I can control access to my data)	Humanities (1) Fine arts (1)	Work with large audio, video, interview data with confidential information
Research collaboration software tools (e.g. Latex, Overleaf, versioning software or other tools that allow secure storage and/or online collaboration between members of research team <i>during the research project</i>)	Humanities (1) School of Business (2)	Work in large teams
Digitization of paper records (lab or field notebooks, paper questionnaires, etc.)	Science (1) School of Business (2)	Working with interview data or secondary data in paper format
Data sharing and access (copyright, patents, licenses, ownership of data, data sharing agreements, etc.)	Science (1) Humanities (1)	
Help in ensuring the confidentiality/anonymity of data on human participants	Science (1)	Working with human subjects
Finding and accessing existing data sources (data scraping, secondary datasets)	Engineering (1)	Looking for data not in researcher's field

Here is a breakdown of workshop interests by faculty

Faculty	Workshops
Engineering	<ul style="list-style-type: none"> • Finding and accessing existing data sources (data scraping, secondary datasets) • Data management activities that will be required by grant funding agencies
Science	<ul style="list-style-type: none"> • Help in ensuring the confidentiality/anonymity of data on human participants • Data sharing and access (copyright, patents, licenses, ownership of data, data sharing agreements, etc.) • Digitization of paper records (lab or field notebooks, paper questionnaires, etc.) • Data documentation and metadata (i.e. data dictionaries, data codebooks containing description of dataset and variables; other supporting documentation) • Data management activities that will be required by grant funding agencies • Data management plan (organization and curation of data throughout research project) • Data management workshop (examples of topics: setting up electronic files, file formats, IP and licensing, de-identification of data)
Humanities	<ul style="list-style-type: none"> • Data sharing and access (copyright, patents, licenses, ownership of data, data sharing agreements, etc.) • Research collaboration software tools (e.g. Latex, Overleaf, versioning software or other tools that allow secure storage and/or online collaboration between members of research team <i>during the research project</i>) • Data archiving (e.g. finding appropriate data repository, understanding how I can control access to my data) • Data documentation and metadata (i.e. data dictionaries, data codebooks containing description of dataset and variables; other supporting documentation) • Data management plan (organization and curation of data throughout research project) • Data visualization (interactive graphics, etc.)

School of Business	<ul style="list-style-type: none"> • Digitization of paper records (lab or field notebooks, paper questionnaires, etc.) • Research collaboration software tools (e.g. Latex, Overleaf, versioning software or other tools that allow secure storage and/or online collaboration between members of research team <i>during the research project</i>) • Data visualization (interactive graphics, etc.)
--------------------	---

Conclusion

The survey and interviews demonstrated that a majority of researchers are interested in adopting better RDM practices, in documenting their data, preserving it and sharing it with colleagues and, in some cases, making it public. Most faculty have a clear awareness of the advantages of performing RDM, both for the researcher producing the data him or herself and for the scientific community at large.

At this moment however, only a small minority of researchers have adopted good data management practices. When documentation exists, it rarely follows accepted standards; when back up procedures are in place, they rarely are systematic or use appropriate safety checks; when data is preserved after the completion of the project, it is often kept on a single hard drive or, at best, on a lab or departmental server; when datasets are shared, this is done informally, often by hand-to-hand or email transfer.

The obstacles to adopting best RDM practices are numerous and have been explored in the literature. The participants in our survey –and in the 2013 survey—have mentioned, among other things:

- the lack of academic incentive for managing and sharing data;
- the scarcity of necessary resources –both human and technological—necessary to curate their datasets;
- the difficulty of dealing with confidentiality and ethics issues
- the desire, for some researchers, to keep a level of control on who will be using their data and for what purposes

It appears clear that the upcoming data management requirement from the Tri-Agency will be a turning point in the attitude and practices of researchers concerning RDM. This is evidenced by the more than 75% of respondents that mention that they would be interested in attending workshops on data activities funded by funding agencies. We can thus expect a fairly rapid change in demand for assistance related to RDM services in Canadian universities whenever those new requirements are enacted.

In order to be ready for that increased demand, we should continue the development of RDM services and maintain a good level of flexibility in the planning and resourcing for those services. We will most probably need to adjust the service offer as the situation evolves quickly over the next few years. A RDM Services Project Charter has been written which describes what is planned for Concordia University.

Appendix 1: Survey Questionnaire

Survey of research data management at Concordia

The aim of this survey is to better understand how Concordia researchers work with their data. The results of this survey and optional follow-up interviews will inform the Concordia Libraries' strategic planning for data services and help PERFORM's Data Science Lab design resources and workshops to meet your needs. In the short term, it will also allow the Data Librarian to provide some assistance to participants.

Participation is voluntary and all published results of this survey will be anonymous. Publications or reports based on the survey or the subsequent interviews will present results in aggregate and will not contain identifying characteristics of respondents.

DEMOGRAPHICS

- Faculty: (Arts & Science, Fine Arts, Engineering, JMSB, Other)
- Department
- Rank: (Lecturer, Assistant Professor, Associate Professor, Professor)
- Research center/laboratory/team (if applicable):

RESEARCH DATA

For the following sections of the survey, think of a research project dataset that **you consider to be your most important**. This should be research data collected or created by you or your team, not a pre-existing or secondary dataset. The dataset may include multiple data files from the same research project. If you cannot choose a most important dataset, simply consider your most recently collected data.

DATA COLLECTION & CURATION

1) Do you collect/create/analyze/store research data?

Research data is defined very broadly to include all primary data collected during your various research projects at Concordia. This might include survey data, experimental data, simulation data, programming code, qualitative data (interview transcripts, field notes, audio or video files), or other information.

- Yes
- No (If no, end survey)

2) How was your research data collection funded? (Choose ALL that apply)

- NSERC
- CIHR
- SSHRC
- CCA
- FQR-S
- FQR-NT
- FQR-SC
- CALQ
- Internal funding
- Private funding
- Other (please specify)

3) How did you collect raw/original data? (Choose ALL that apply)

- Paper and pencil (questionnaires, laboratory notebooks, recording sheets)
- Collect specimens or physical samples (water, tissue, cells, blood, soil, etc.)
- Direct electronic input (iPad/digital forms/online)
- Record images, observations, events
- Other - Specify: _____

DATA MANAGEMENT

4) Who is responsible for managing your data? (Choose ALL that apply)

- Yourself
- Research Coordinator/Project manager
- Graduate student
- Designated person in the research group
- Research assistant
- External Research Collaborator/Project Partner outside of Concordia

- IT staff in the department
- Nobody
- Other (please specify)

5) How did you convert raw/original data into a structured electronic form/database? (Choose ALL that apply)

- Data are collected in structured electronic form
- Manual data entry/transcription
- Optical Character Recognition scan to derive data from paper forms (OCR)
- Scan of paper form (PDF)
- I don't put data into structured electronic form/database
- Other - Specify: _____

6) What type of data was collected or created for this specific project? (Choose ALL that apply)

- Text (e.g. TXT, DOC, PDF, RTF, HTML, XML)
- Numerical or statistical (e.g. CSV, MAT, XLS, SPSS, R, etc.)
- Multimedia, including images, audio, or video (e.g. JPEG, TIFF, MPEG, QuickTime, Bitmap)
- Databases (e.g. MS Access, Oracle)
- Instrument specific (e.g. LC-MS, ECG, QPCR)
- 3D Imaging (e.g., LORIS, MRI)
- Geospatial (e.g. ArcGIS, MapInfo, QGIS, WorldMap)
- Models (e.g. 3D, statistical, similitude, macroeconomic, causal)
- Programming code (e.g. Java, C, Perl, Python, Ruby, PHP, JSON, Dplyr, D3, HTML)
- Other Please specify: _____

7) What type of supporting documentation do you keep on your data? (Choose ALL that apply)

Supporting documentation allows someone outside your research group to understand and use the research data or to replicate the methodologies that produce the data. This documentation can be retained in the same file, folder or document as the research data.

- Data dictionary/Coding manual

- Description of methodology / protocol
- Questionnaire or interview text
- Instrument metadata (manufacture, model, parameters, calibration)
- Programming code (to replicate analysis)
- Additional metadata (e.g., technical, preservation, copyright, access restriction)
- Data citations (e.g., source of merged data with your original data)
- Other - Specify: _____
- I do not keep any data documentation

8) What is the approximate total size of this dataset?

- Less than 10GB (Small USB stick is 8GB)
- Between 10 and 100GB (Portable Pocket Drive)
- Between 100GB and 4TB (Large External Hard Drive)
- More than 4TB (Server)
- I don't know

9) DURING the research project and data collection, where do you store your working dataset/electronic files? (Choose ALL that apply)?

- Cloud storage (Dropbox, Google Drive, MS OneDrive, etc.)
- Computer hard drive
- External hard drive/Portable storage/Flash drive (USB)
- Hard drive of the instrument which generates the data
- Department or lab server
- IITS server
- CDs/DVDs/Tape
- Data Repository (ICPSR, GenBank, PubChem, figshare, DRYAD, GitHub, LORIS, Harvard's Dataverse, Scholars Portal's Dataverse, etc.)
- Physical hardcopy retained (in boxes, cabinets, etc.)
- Other (please specify)

10) How do you secure your electronic data? (Choose ALL that apply)

- Password protected
- Data is de-identified (i.e., only anonymous data stored)
- Data is encrypted
- Data is not secured
- Other. Specify _____

11) How often do you back up your working data?

- Daily
- Weekly
- Monthly
- Ad Hoc
- Never
- Other (please specify)

12) Where do you store your back up copies (additional copies of your working datasets kept in case of data loss?) (Choose ALL that apply)

- Cloud storage (Dropbox, Google Drive, MS OneDrive, etc.)
- Computer hard drive
- External hard drive/Potable storage/Flash drive (USB)
- Hard drive of the instrument which generates the data
- Department or lab server
- IITS server
- CDs/DVDs/Tape
- I don't store backup copies
- Other (please specify)

13) What software or programs do you use when collaborating with others? (Choose ALL that apply)

- Word Track Changes
- Google Documents

- Shared Folders (DropBox)
- Overleaf, Authorea, DocEar (LaTeX)
- GitHub
- Referencing (Refworks, BibTex, EndNote, Papers)
- None
- Other - Specify: _____

14) How do you deal with different versions of your working data files?

- I save multiple files and update the files with different names/naming convention
- This is done automatically (e.g., Dropbox)
- I use a versioning control program (NAME: _____)
- I have not used versioning control, but I am keen to learn more.
- I only use one dataset version.
- Other (please specify)

DATA STORAGE AND ARCHIVING AT THE END OF PROJECT

15) Where do you store your dataset at the conclusion of your research project? (Choose ALL that apply)

- Cloud storage (Dropbox, Google Drive, MS OneDrive, etc.)
- Computer hard drive
- External hard drive/Portable storage/Flash drive
- Hard drive of the instrument which generates the data
- Department or lab server
- IITS server
- CDs/DVDs/Tape
- Physical hardcopy retained (in boxes, cabinets, etc.)
- I do not keep a copy of my dataset [**skip to 17**]
- Other

16) How long do you retain your dataset after the completion of your research project?

- 1 year
- 5 years
- 10 years
- Indefinitely
- I do not know
- Other - Specify: _____

17) After your research project is completed –the data has been cleaned and analyzed– in what data repository would you consider archiving this dataset? (Choose ALL that apply)

- SPECTRUM (Concordia’s institutional repository)
- General or discipline-specific repository (ICPSR, Protein Data Bank, Cambridge Structural Database, GitHub, Dryad, Figshare, Harvard’s Dataverse, Scholars Portal’s Dataverse, etc.)
- Other – Specify

- I do not know about data repositories, but I am keen to learn more.
- Not interested in archiving [**skip to 19**]

18) How long do you think your dataset should be preserved in a data repository after the completion of your project?

- 1 year
- 5 years
- 10 years
- Indefinitely
- I do not know
- Other - Specify: _____

19) Are you aware of any discipline-specific data repositories in your area of research?

- Yes
Please list them. [Free text]
- No

20) How do you share your research data with others? (Choose ALL that apply)

- Physically hand over data (USB, CD, Portable laser drive)
- Email data files
- Link to data on academic website
- Data repository, open-access (anonymous dataset available to public)
- Data repository, controlled access (approved researchers only/data sharing agreements)
- I have not been sharing my data so far, but would like to learn more about different options
- I prefer not to share my data
- Other (please specify)

WORKING WITH DATA: STATISTICS, PROGRAMMING, & VISUALIZATION

21) What software do you use for statistical analysis? (Choose ALL that apply)

- SAS
- R
- SPSS
- Stata
- Matlab
- Excel
- NVivo / Qualitative software
- Other - Specify: _____
- I do not conduct statistical analyses

22) Where do you go for help with unfamiliar statistical procedures? (What resources are available to you when you have a challenging statistical issue/problem?) (Choose ALL that apply)

- Colleagues within department
- Faculty outside of department
- Online resources
- Specialized research centers

- Google/YouTube
- Workshop - external
- QICSS Consultation
- I do not need help with statistics
- Not applicable
- Other (please specify)

23) Do you use programming or write code for any of the following research data activities? (Choose ALL that apply)

- Data collection
- Data manipulation and management
- Statistical and computational analysis (including simulations)
- Graphics and manuscript preparation
- Webpage creation
- Other. Specify _____
- I have no/little experience programming, but I am keen to learn more.
- Not applicable

24) In what languages do you program? (Choose ALL that apply)

- Python
- Perl
- Java
- C++
- HTML/CSS
- Not applicable
- Other. Specify _____

25) What tools do you use for creating graphs, figures, or infographics? (Choose ALL that apply)

- Excel

- Statistical Software (SPSS, Stata, R)
- Adobe Creative Suite
- Java Script (D3)
- Other - Specify: _____
- Not applicable

26) Thinking about peer-review journals in your discipline, or other respected sources for research dissemination, what is your experience with the following: (Choose ALL that apply)

- I have read articles with interactive graphics
- I have submitted manuscripts with interactive graphics
- I have submitted manuscripts to journals that required data files/statistical code be made accessible
- I avoid journals that have an open-access requirement to link my data files

DATA WORKSHOPS/ASSISTANCE

27) For each of the following, please indicate your level of interest in data management assistance at Concordia or training/development opportunities (workshops/resources):

	Very interested	Somewhat interested	Not interested	Not applicable
Data management workshop (examples of topics: setting up electronic files, file formats, IP and licensing, de-identification of data)				
Data visualization (interactive graphics, etc.)				
Data management plan (organization and curation of data throughout research project)				
Data management activities that will be required by grant funding agencies				
Data documentation and metadata (i.e. data dictionaries, data codebooks containing description of dataset and variables; other supporting documentation)				

Data archiving (e.g. finding appropriate data repository, understanding how I can control access to my data)				
Research collaboration software tools (e.g. Latex, Overleaf, versioning software or other tools that allow secure storage and/or online collaboration between members of research team <i>during the research project</i>)				
Digitization of paper records (lab or field notebooks, paper questionnaires, etc.)				
Data sharing and access (copyright, patents, licenses, ownership of data, data sharing agreements, etc.)				
Help in ensuring the confidentiality/anonymity of data on human participants				
Finding and accessing existing data sources (data scraping, secondary datasets)				
Other (please specify)				

28) Are you aware of any resource or project in your department or faculty aimed at facilitating research data management?

- Yes. Please describe.
- No

Thank you for completing this survey! Your participation is sincerely appreciated. If you would like to receive information about planned data science activities or workshops, please provide your contact information below. Also indicate if you would be willing to be contacted to participate in a follow-up interview.

- Name:
- Email:
- Phone:
- Department:
- Willing to participate in follow up interview? (Y/N)
- Keep me informed of workshops or data science activities for me or my students? (Y/N)

If you have any questions or comments about this survey, please use the text box below or contact one of the researchers:

Alex Guindon (alex.guindon@concordia.ca)

Danielle Dennie (danielle.dennie@concordia.ca)

Jennifer McGrath PERFORM Data Science (data.science@concordia.ca)

Appendix 2: Research Data Management Interview protocols

Social Sciences, Humanities, Business, Fine Arts

This interview is part of the Survey of Concordia University Research Data Management project conducted by Alex Guindon, GIS and Data Services Librarian

NOTE: The content of this interview form is adapted from McGowan, T. & Gibbs, T. A. (2009) Southampton Data Survey: Our Experiences & Lessons Learned [unpublished]. University of Southampton: UK.

Introduction

My name is [Alex Guindon or Danielle Dennie]; I am the *Data and GIS services librarian* [or reference/subject librarian] at Concordia. I am grateful that you accepted to meet with me to discuss the management of your research data. The interview should take approximately one hour. I would like to record our conversation if you do not object. What you say during our conversation will be confidential. Quotes appearing in any report or publication based on this interview will also be anonymous unless you give me your express consent to use your name.

Structure of the interview

This interview is a follow up to the survey questionnaire that you have already completed and in which you identified your most important data asset. As was explained in the email you received before this meeting, there will be three themes to this discussion: 1) a detailed description of your research data and some context about your research area; 2) the issue of archiving and sharing your dataset; 3) your needs in terms of data management assistance. This exercise will allow us to answer some of your questions and to direct you to appropriate resources for data curation. I will also present a report to the Libraries' administration that will inform our strategic planning for data curation.

The intent of this meeting is to have a discussion about your data management practice rather than a formal question and answer session. Please do not hesitate to ask questions or to address issues that are important to you even if they do not correspond to a specific interview question.

Discussion outline

A. Description of dataset and research context

1. Could you please tell us a bit about your area of research?
2. Could you please give us some details about what you would consider your most important dataset? (Try to obtain information for as many of the elements in the table below as possible)
What is the name of this dataset, or the project that is generating this data?

Tick	Metadata Heading	Description
	Author	Person, group or organisation responsible for the intellectual content of the dataset
	Owner(s)	Current legal owner(s) of the dataset
	Source	The source(s) of the information found in the data asset
	Purpose	Reason why the asset was created, intended user communities or source of funding / original project title
	Title	Official name of the data asset, with additional or alternative titles or acronyms if they exist
	Description	A description of the information contained in the data asset and its spatial, temporal or subject coverage
	Subject	Data topics and keywords describing the subject matter of the data
	Geographical coverage	The countries, regions, cities etc covered in the data
	Time period covered	The date (or date range) covered by the data
	Date of collection	The date (or date range) on which the data was collected
	Sample size & description	The number of individuals surveyed and characteristics
	Current location	Path or www address where the data can be found
	Format	Physical formats of dataset, including file format information
	Size	Size of the data in Mb/Gb
	Documentation available	Documentation that is available (e.g. user manuals, codebooks), including references to its location

3. Where would you consider storing your data while still working on it?

B. Preservation of dataset

- a. Do you think preservation of this dataset is important? If so why?
- b. How long do you think the dataset should be preserved?
- c. Where would you consider archiving your data and what would be the characteristics of such an archive (for example, type of access, privacy, size, format, length of preservation)?

C. Sharing of dataset

If respondent indicated interest in sharing data (Q20 of the survey):

- a. In the survey, you mentioned that you would like to share this dataset. Can you tell me a bit more about who would be interested in reusing this data?
- b. Are there any confidentiality issues associated with your data?
- c. (if applicable)
Did your participants sign a consent form? Did the form mention the possibility of data reuse (sharing data with other researchers)?
- d. Discuss the type of access preferred by the faculty (based on answer to Q20 of the survey)

If respondent did not indicate interest in sharing data

- e. In the survey, you mentioned that you preferred not to share your data. Would you like to discuss this further?
Do you believe that the data is too sensitive or that there are confidentiality issues that cannot be addressed?
Do you consider that your data could be reused for secondary analysis? (i.e. asking different research questions)?
- f. Mention the various types of access (open, restricted, enclave). Would any of these access management methods alleviate your concerns?

-
- g. How long should you dataset be preserved?
 - h. Are there any special archiving issues associated with your data? Examples: format migration, physical space, dynamic database (still being updated), need to keep several versions, etc?

D. Requirements for services and assistance in data management

In the survey (Q27), you indicated that you would like to see this type xxx of data management assistance being offered at Concordia. Can you tell me more about explain what your expectations would be in regard to this service?

Is there any other type of data management help you would like to see offered at Concordia?

E. Interview wrap-up

- a. Is there anything else you would like to discuss regarding data management?
- b. Do you have any questions about this research project?

Thank you

Thank you for your participation in this project. Your answers, both to the online survey and to this interview, will be very useful as we plan our data services at Concordia. I will communicate with you regarding data archiving and data sharing if you wish to go in that direction. Do not hesitate to contact me should you have any other thoughts or questions regarding data management.

Science & Engineering

This interview is part of the Survey of Concordia University Research Data Management project conducted by Alex Guindon, GIS and Data Services Librarian

NOTE: The content of this interview form is adapted from McGowan, T. & Gibbs, T. A. (2009) Southampton Data Survey: Our Experiences & Lessons Learned [unpublished]. University of Southampton: UK.

Introduction

My name is [Alex Guindon or Danielle Dennie]; I am the *Data and GIS services librarian* [or reference/subject librarian] at Concordia. I am grateful that you accepted to meet with me to discuss the management of your research data. The interview should take approximately one hour. I would like to record our conversation if you do not object. What you say during our conversation will be confidential. Quotes appearing in any report or publication based on this interview will also be anonymous unless you give me your express consent to use your name.

Structure of the interview

This interview is a follow up to the survey questionnaire that you have already completed and in which you identified your most important data asset. As was explained in the email you received before this meeting, there will be three themes to this discussion: 1) a detailed description of your research data and some context about your research area; 2) the issue of archiving and sharing your dataset; 3) your needs in terms of data management assistance. This exercise will allow us to answer some of your questions and to direct you to appropriate resources for data curation. I will also present a report to the Libraries' administration that will inform our strategic planning for data curation.

The intent of this meeting is to have a discussion about your data management practice rather than a formal question and answer session. Please do not hesitate to ask questions or to address issues that are important to you even if they do not correspond to a specific interview question.

Discussion outline

F. Description of dataset and research context

4. Could you please tell us a bit about your area of research?
5. Could you please give us some details about what you would consider your most important dataset? (Try to obtain information for as many of the elements in the table below as possible)
What is the name of this dataset, or the project that is generating this data?

Tick	Metadata Heading	Description
	Author	Person, group or organisation responsible for the intellectual content of the dataset
	Owner(s)	Current legal owner(s) of the dataset
	Source	The source(s) of the information found in the data asset
	Purpose	Reason why the asset was created, intended user communities or source of funding / original project title
	Title	Official name of the data asset, with additional or alternative titles or acronyms if they exist
	Description	A description of the information contained in the data asset and its spatial, temporal or subject coverage
	Subject	Data topics and keywords describing the subject matter of the data
	Geographical coverage	The countries, regions, cities, etc., covered in the data
	Time period covered	The date (or date range) covered by the data
	Date of collection	The date (or date range) on which the data was collected
	Sample description	The spatial, biotic or abiotic sample used, its characteristics, and parameters measured
	Current location	Path or www address where the data can be found
	Format	Physical formats of dataset, including file format information
	Size	Size of the data in Mb/Gb
	Documentation available	Documentation that is available (e.g. user manuals, codebooks), including references to its location

6. Where would you consider storing your data while still working on it?

G. Preservation of dataset

- a. Do you think preservation of this dataset is important? If so why?
- b. How long do you think the dataset should be preserved?
- c. Where would you consider archiving your data and what would be the characteristics of such an archive (for example, type of access, privacy, size, format, length of preservation)?

H. Sharing of dataset

If respondent indicated interest in sharing data (Q20 of the survey):

- a. In the survey, you mentioned that you would like to share this dataset. Can you tell me a bit more about who would be interested in reusing this data?
- b. Are there any confidentiality issues associated with your data?
- c. (if applicable)
Did your participants sign a consent form? Did the form mention the possibility of data reuse (sharing data with other researchers)?
- d. Discuss the type of access preferred by the faculty (based on answer to Q20 of the survey)

If respondent did not indicate interest in sharing data

- e. In the survey, you mentioned that you preferred not to share your data. Would you like to discuss this further?
Do you believe that the data is too sensitive or that there are confidentiality issues that cannot be addressed?
Do you consider that your data could be reused for secondary analysis? (i.e. asking different research questions)?
- f. Mention the various types of access (open, restricted, enclave). Would any of these access management methods alleviate your concerns?

-
- g. How long should your dataset be preserved?
 - h. Are there any special archiving issues associated with your data? Examples: format migration, physical space, dynamic database (still being updated), need to keep several versions, etc?

I. Requirements for services and assistance in data management

In the survey (Q27), you indicated that you would like to see this type xxx of data management assistance being offered at Concordia. Can you tell me more about explain what your expectations would be in regard to this service?

Is there any other type of data management help you would like to see offered at Concordia?

J. Interview wrap-up

- a. Is there anything else you would like to discuss regarding data management?
- b. Do you have any questions about this research project?

Thank you

Thank you for your participation in this project. Your answers, both to the online survey and to this interview, will be very useful as we plan our data services at Concordia. I will communicate with you regarding data archiving and data sharing if you wish to go in that direction. Do not hesitate to contact me should you have any other thoughts or questions regarding data management.