acrl 2023

DATA SERVICES IN A POST PANDEMIC ENVIRONMENT:

What do Graduate Students and Postdocs Need?

Jennifer Hart, Debra A. Werner, and Aditi Goyal*

INTRODUCTION

The COVID-19 pandemic precipitated a shutdown of physical workspaces on university campuses all over North American. This event had a significant effect on the work processes of graduate students and postdoctoral researchers in the sciences. In many cases they were unable to physically access their laboratories and other university facilities for a period of time. As a result, many of their workflows were disrupted, including those for data collection, processing, and management. These disruptions exposed limitations in laboratories' policies and procedures, the organization of research facilities, and resources and tools available to students and postdocs for their work.

The University of Chicago is an R1 research institution with over 17,000 students of which nearly 10,500 are graduate students.[1] The sciences are divided into two divisions, the Biological Sciences Division (BSD) and the Physical Sciences Division (PSD), and the Pritzker School of Molecular Engineering. The BSD is comprised of ten departments in the biological sciences: Cancer Research, Biochemistry and Molecular Biology, Ecology and Evolution, Human Genetics, Microbiology, Molecular Genetics and Cell Biology, Neurobiology, Organismal Biology and Anatomy, Pharmacological and Physiological Sciences, and Public Health Sciences.[2] In addition, the University Medical Center's clinical programs are part of the BSD. The Physical Sciences Division is composed of seven departments: Astronomy and Astrophysics, Chemistry, Computer Science, Geophysical Sciences, Mathematics, Physics, and Statistics as well as a number of interdisciplinary research institutes and centers.[3] Both divisions provide access to many Core Facilities and other resources that offer students and postdocs the use of instruments, technologies and services for their work beyond what is available in their laboratory.[4]

In winter and spring of 2021, a physical sciences librarian and a biomedical librarian conducted focus groups of graduate and postdoctoral students in the BSD and PSD to better understand their research data management practices and what support they needed with data management. The questions for the focus groups were written before the COVID pandemic. However, because of the timing of the sessions, conducted a year into the pandemic after many of the participants' work had been affected by the physical shutdown of campus facilities, many of their responses to questions reflected the conditions they worked under during the shutdown period and what problems they encountered working with data during this time. They were also

* *Jennifer Hart, Computer Science, Mathematics, Physics and Statistics Librarian, University of Chicago, hartj@uchicago.edu; Debra A. Werner, Biomedical Librarian, University of Chicago, dwerner@uchicago.edu; Aditi Goyal, Graduate Fellow, University of Chicago, aditig@uchicago.edu*

FIGURE 1

North Science Quadrangle at the University of Chicago

able to reflect on how their work processes had changed as a result of the shutdown and how they preferred to work in a post pandemic environment.

METHODS

This study was comprised of four focus groups that were conducted via Zoom with a total of 11 participants. It was reviewed and exempted by the Institutional Review Board at the University of Chicago.

Participants in the study were grouped by similar research activities to the extent possible. The focus groups took the form of semi-structured interviews with a script of prepared questions. They were asked questions about the kind of data they collected and how they collected it, who in their lab was responsible for managing the data, and how their data was managed and stored. They were also asked about the data services they had available to them as well as what data services they would like to use that they did not presently have access to. A full list of questions can be found in the appendix.

Interviews were recorded and then transcribed. Two librarians and a data management fellow coded transcripts in Dedoose and performed a thematic analysis to inform their conclusions.

FINDINGS

Overall, participants encountered a number of issues managing their data. They reported problems both with shared practices within their labs and with their own data management practices.

If data practices were established within a lab, they were often developed and managed by the lab's Principal Investigator (PI) or someone else specifically tasked with that role. In some cases, a graduate student or postdoc would take the initiative to create shared practices. When shared practices were lacking, however, participants reported they often had trouble accessing needed data. Specifically, they found information commonly used within labs, such as protocols, were often not shared, or, in some cases not written down at all. They also had trouble accessing and working with previous students' data due to inconsistencies in how data was managed and shared in their lab. Previous students would sometimes take their data with them when they left, or would leave it in a form that was not easy to access or interpret.

With regard to their personal data management practices, participants reported having trouble with organizing and naming their files. They also had issues with processing, sharing and storing large data files. In addition, they had limited knowledge of metadata practices and some didn't employ common practices such as readme files. Without this metadata, they encountered problems. For example, one mentioned having to interpret abbreviated names on data samples that referred to a record on a spreadsheet. However, the spreadsheets were poorly organized and there was no metadata attached to them. Consequently, it was difficult to match the sample and record.

In addition, specific themes emerged regarding issues participants had resulting from the pandemic. One was their difficulty accessing data remotely from laboratories, Core Facilities and other resources on campus due to university, division, and laboratory policies. Another issue was their lack of access to needed software and hardware to process and analyze their data because of limitations in licensing and funding, and the prohibitive expense for them to personally license needed software. A third was inequities and inconsistencies in the policies of departments and their funding of resources that affected participants' ability to manage their data.

Policies and Security Protocols in Accessing Data

One of the most common problems participants had during the shutdown was they had difficulty accessing the data they needed for their work. This was usually due to the policies and procedures in place in their laboratories and in other facilities they used on campus. Many laboratory and facility policies had been established with the assumption students and postdocs would be able to work on site. There were also policies in place to address security concerns about data and to prevent laboratory and facility produced data from being hacked. Some of these security protocols were also required by data management plans within the labs. As a result of all these various policies, labs and facilities sometimes imposed strict rules as to where data could be saved and how it could be accessed. In many cases, these limits made it more difficult for students and postdocs to access needed data remotely and sometimes impossible.

Participants reported that some laboratories computers were not connected to the internet at all. This was often in order to maintain the strictest level of security on these computers. In some cases, the computers were connected to laboratory instruments and used to collect and store data from these instruments. Data transfer was only possible from them via USB. When students or postdocs no longer had physical access to the laboratory they had no way to retrieve this data. They also noted using a USB for data transfer had its own drawbacks as it could spread viruses that might compromise or destroy files.

In other instances, data was stored on computers that were connected to the internet, but the data could not be uploaded to the cloud. Remote access to these computers was allowed, but students and postdoc had difficulties remoting in. If the data they needed was stored on a single computer only one person at a time could access their data remotely. This limited how easily and frequently they could work with needed data.

Before the shutdown, participants had access to other computers in the lab with the correct security software installed and were allowed to move data over to these computers both from other lab computers and from the Core Facilities they used to collect data. But in a remote environment they sometimes could no longer do this. As a result, they also had to process this data remotely. They often needed access to computers for a long period of time to work with large data sets.

Some participants reported they were unable to remote in at all due to the incompatibility of the remote desktop software they used and what the lab was willing to have installed on their computers due to security concerns.

Participants also mentioned having to be very careful with data if it was only available on one place as they feared damaging data or files others also needed to access.

Collaboration Issues

In some cases, participants had developed relationships with labs at other universities or research institutions and shared use of each institutions' facilities and instruments for data collection. Difficulties they had accessing data due to security protocols also extended to their work with these collaborators. They had issues both sharing their data with them and working with data collaborators were able to provide.

Participants reported that when they needed to collect and work with data provided by their collaborators, security policies these institutions had in place affected how they could access needed data. In some cases, they needed special permission from the institution to access data remotely. They sometimes had to resort to processing data remotely due to security concerns with transferring data.

In cases where they shared University of Chicago generated data, participants reported other concerns. Some labs and facilities were able to upload data to a secure cloud storage such as Box, but their collaborators were not able to access it due to security concerns. Lacking an alternative secure method, they were also forced to process the data remotely on the computer or server on which it was originally gathered.

The difficulty participants had during the pandemic accessing and working with data also led them to create workarounds. If they couldn't download or transfer data, some would take images or screenshots of the relevant data or download the processed data as a CSV file and email these files to themselves. They thought some of these workarounds might be an insecure way to transfer data, but said was what they had available to them.

Software & Hardware Issues

Participants also reported problems accessing the software they needed during the pandemic. These issues then affected their work processes. Often they lacked access to the most useful software for their work, usually because it was not provided by the lab and participants could not afford to purchase it themselves. They often used free, open source, or low-cost alternatives in place of preferred, but expensive, software. However, in many cases, their substitute was inferior in significant ways and made their work more difficult.

In some cases, the lab provided a limited license to the best software available for their work, but it was only installed on a desktop or desktops inside the lab. During the shutdown, they either couldn't access that software at all or had to remote in to process needed data with it. In other instances, they had access to an outdated version of the software and could work with it from home. However, these older versions might lack features they needed. One participant reported they could not transfer files using an outdated version, and instead had to take screenshots in order to save the information.

In cases where they did not have access to an outdated or free version of the needed software, they would often substitute a similar software, usually an open-source option.

An example of the kinds of problems they encountered during the shutdown was their use of remote desktop software. Most students and postdocs needed to employ some kind of remote desktop software to access data in their lab, but they often did not have software available to them that was optimal for their needs. A number of them reported having problems with the software they did use. In some cases, due to security concerns, the lab would not install the same software on lab computers that participants had. They then could not access lab computers at all. In other instances, because their home computer was a Mac and they were remoting into a PC, they required special software to work with applications in the different operating system. In addition, outdated software could have other issues. One participant noted that their software displayed too small of a resolution on their home computer, making work difficult.

Participants did mention that their access to some softwares was improved during the pandemic. Some software licenses that had, in the past, only been available to them on an individual lab desktop were made temporarily available for free to students during the shutdown so they could work with it from home. One example of this was the Adobe Creative Suite. Usually, this special access was discontinued after the shutdown. Students expressed a desire for this access to be permanent as it made their work so much easier.

Overall, there were a number of softwares desired by participants, but were cost prohibitive for them to purchase individually. The one most commonly named for image manipulation was the Adobe Creative Suite.[5] Origin and Prism were also commonly desired for graphing and data analysis.[6] They were also interested in electronic lab notebooks, citation managers and remote desktop software, but they did not have a common preferred product for those softwares.

In addition to their lack of needed software, participants reported issues procuring the hardware they needed to work effectively at home. A number of them mentioned using a laptop for their work, but some could not afford to purchase one themselves.

Even if they were able to procure a personal laptop, it could be inadequate for their needs. Files produced by Core Facilities, for example, were often large, and participants had issues transferring them to their computers. They often didn't have the storage space or processing power on their personal computer to work effectively with large data sets.

They also lacked basic hardware and infrastructure at home for their work such as a second monitor to work with large datasets more effectively and a fast and reliable internet service.

FIGURE 2

Student at Work

Inequities in Access to Hardware and Software

There were also inequities in the availability of needed software for students and postdocs. Some labs provided group licenses, where others did not. When a license to needed software was not available, some participants had other means of access such as a license from a previous institution they were affiliated with. Others either had to purchase the software themselves, or if they could not afford it, rely on the open-source options available to them.

With regard to laptops specifically, some departments allowed students to use funds available to them to purchase a laptop. In other departments, students reported that it was unclear if they would be taxed on these funds if they employed the laptop for personal use. Because of this lack of clarity, they avoided using the funds to purchase a needed laptop and had to either pay for one themselves or work without it.

DISCUSSION & RECOMMENDATIONS

The shutdown exposed problems with how students and postdocs are expected to access and work with data and the tools they have available for them to do so. In the aftermath of the pandemic, students and postdocs have adapted to working in a more remote environment, and consequently, many of the issues they experienced during that time persist.

Most participants reported they collect large datasets for their work, and spent considerable amounts of time processing the data. They would like to be able to do more of their work with data offsite and to have the means to process data on their own computer. Labs could facilitate this by providing comprehensive remote access to data through cloud storage, remote desktop software or by other means while still adhering to their security protocols. Students and postdocs also would benefit from improved ways to share large data sets with collaborating institutions. Even when cloud storage was available to them, it was not necessarily accessible to their collaborators. They expressed an interest in better pipelines for sharing data between institutions.

Participants also made clear they highly value access to the specialized software necessary for their work. Lacking this access was a significant barrier for them, both limiting where and how they could work. It also limited how efficiently and securely they could work with data. Broader licenses to softwares, when possible, would be of great benefit to them.

Students and postdocs clearly need help with basic data management skills. Few were aware of any data management plan within their lab, and only some labs had developed any kind of shared file structures, file naming conventions, or standards for keeping and organizing protocols. Participants also reported they had difficulty managing their own data and were unsure of the best way to do it. Learning basic data management principles such as naming conventions, file organization, metadata standards and effective data preservation would be very valuable to them. This type of training would improve efficiencies for both the participants and the labs and reduce problems such as the case of one participant who reported abandoning an inherited project because, due to insufficient documentation, they were unable to recreate their predecessor's technique. Librarians could offer this kind of data management skills training.

Universities should make common licenses for software available to students and postdocs. This would eliminate the need for time intensive workarounds that don't always adhere to security policies. Common licenses could also help facilitate better instruction on data management issues. If large cohorts of students were using the same software, librarians would be able to teach data management skills tailored to its specific capabilities.

Many universities and institutions are making concerted efforts to address equity issues. Providing institution-wide software licenses would give fair access to those who cannot afford to purchase needed software themselves or do not have other means to acquire it. In addition, improving policies to provide equitable funding for laptops would be valuable, as they are essential for students' and postdocs' work. A more equitable environment would help level the playing field for all students.

APPENDIX: FOCUS GROUP QUESTIONS

1) Briefly describe a project you have worked on.
    1a) What types of data are collected for the project?

2) Does this project have a written data management plan?
    2a) Who developed the data management plan for this project?
    2b) Who has responsibility for managing your data?
    2c) Which tools do you use to manage your data?
    2d) Which file formats do you use?
    2e) How much file storage do you typically need for your data?
    2f) Do you use a data repository to manage your data? If so, which one(s)?

3) Who is responsible for the creation of metadata and file schemes for your research?

4) Will data collected for or generated by this project be reused? And if so, by whom?
    4a) If so, how will the data be made available?
    4b) If not, do you think there might be value to others in reusing your data?

5) What data management challenges have you experienced?
    5a) Please describe any data management failures you have experienced.

6) Which data services have you used at the University
    6a) Have you used external data services, including those provided by a collaborator's institution? If so, please describe them.

7) Which, if any, data services do you wish you had available to you?

8) What else should we know that we haven't asked you about?

NOTES

[1] "Facts and Figures," The University of Chicago, 2023, https://www.uchicago.edu/who-we-are/facts-and-figures
[2] "About," Biological Sciences Division, 2023, https://biologicalsciences.uchicago.edu/about
[3] "About Us," Physical Sciences, 2023, https://physicalsciences.uchicago.edu/about/
[4] "Core Facilities," Biological Sciences Division, 2023, https://biologicalsciences.uchicago.edu/resources/osrf-core-facilities; "Facilities & Services," Physical Sciences, 2023, https://physicalsciences.uchicago.edu/resources/facilities-services/
[5] "Adobe Creative Cloud," Adobe, 2023, https://www.adobe.com/creativecloud.html
[6] "OriginPro 2023," OriginLab, 2023, https://www.originlab.com/; "Prism," GraphPad by Dotmatrix, 2023, https://www.graphpad.com/scientific-software/prism/

BIBLIOGRAPHY

Cox, Andrew M., and Stephen Pinfield. "Research data management and libraries: Current activities and future priorities." *Journal of Librarianship and Information Science* 46 (4) (2013): 299-316. https://doi.org/10.1177/0961000613492542.

Joo, S., and C. Peters. "User needs assessment for research data services in a research university." *Journal of Librarianship and Information Science* 52 (3) (2020): 633-646. https://doi.org/10.1177/0961000619856073.

North Carolina State University Libraries. "Defining Research Data." Accessed February 15, 2023. https://www.lib.ncsu.edu/do/data-management/defining-research-data.

Pasek, Judith E., and Jennifer Mayer. "Education Needs in Research Data Management for Science-Based Disciplines: Self-Assessment Surveys of Graduate Students and Faculty at Two Public Universities." *Issues in Science and Technology Librarianship* (92) (Fall 2019). https://doi.org/10.29173/istl12.

Pouchard, Line, and Marianne Stowell Bracke. 2016. "An Analysis of Selected Data Practices: A Case Study of the Purdue College of Agriculture." *Issues in Science and Technology Librarianship* (85) (Fall 2016). https://doi.org/10.5062/F4057CX4.

Valentino, Maura, and Michael Boock. "Data Management Services in Academic Libraries: A case study at Oregon State University." *Practical Academic Librarianship: The International Journal of the SLA Academic Division* 5 (2): 77-91 (2015). https://pal-ojs-tamu.tdl.org/pal/article/view/7001.

Weller, Travis, and Amalia Monroe-Gulick. "Differences in the Data Practices, Challenges, and Future Needs of Graduate Students and Faculty Members." *Journal of eScience Librarianship* 4 (1) (2015). https://doi.org/10.7191/jeslib.2015.1070.

Wiley, Christie, and Erin E. Kerby. "Managing Research Data: Graduate Student and Postdoctoral Researcher Perspectives." *Issues in Science and Technology Librarianship* (89) (Spring 2018). https://doi.org/10.29173/istl1725.