Sarah Webb Holsapple. Data sharing for Master Theses: Survey and Recommendations. A Master's Project for the M.S. in I.S. degree. July, 2021. 53 pages. Advisor: Brad Hemminger

As new practices become established procedures, we must amend policies and guidelines to expand to fit these new norms. Data sharing is now becoming an expectation in scientific research. This paper will survey the landscape of data sharing policy from funders, publishers, and universities and will take from those policies to suggest changes to the UNC SILS master's paper/project policy to include a data sharing component, namely, a data management plan and a data availability statement. In order to assist students following the new policy and sharing data if desired, this project will create a libguide outlining best practices for data sharing and a step by step process to share data at UNC SILS.

Headings:

Information sharing

Open data movement

Data curation

DATA SHARING FOR MASTER THESES: SURVEY AND RECOMMENDATIONS

by
Sarah Webb Holsapple

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

July, 2021

Approved by

_____

Brad Hemminger

1 Introduction

In his landmark 1945 article in the *Atlantic*, "As We May Think," Vannavar Bush predicted an explosion of information and data and outlined a new way to store it all.

> A record if it is to be useful to science, must be continuously extended, it must be stored, and above all it must be consulted. Today we make the record conventionally by writing and photography, followed by printing; but we also record on film, on wax disks, and on magnetic wires (Bush, 1945).

Now we are able to collect and store endless amounts of data on computers. Once all that data is collected and stored, the next question is – can it be consulted? Only if it is shared openly. The push towards open access and open data seeks to make the data and resulting publications accessible to all, data sharing is a part of that and is defined generally as "the practice of making data used for scholarly research available to other investigators" (Wikipedia, 2021).

This project aims to explore the landscape of data sharing policy in the US, outline a set of best practices for data sharing, create a set of guidelines to apply those practices to master's paper data/digital objects at UNC SILS, and propose to UNC SILS that master's paper requirements include a data sharing requirement.

The literature review will answer the question – Why should UNC SILS students share their data from their masters papers. The project itself will answer the question – How can UNC SILS students share their data.

1.1 Background

As I sat down to consider what I would write my paper about, I looked through papers in the Carolina Digital Repository and had some thoughts about the magnitude of the data that is created for these papers, and the work that goes into collecting and processing that data. Once that work is done and the paper is completed that data never really sees the light of day. Not only would this proposal bring that data out, but would open up whole new worlds of inquiry, examining old datasets to see what's changed, aggregating datasets to come up with new lines of research, and potentially using the plans themselves as data.

The Carolina Digital Repository currently hosts 2,196 (as of May 19, 2021) UNC SILS student master's papers, presumably a decent number of those papers involved the creation of new data, either through survey data collection, content analysis, interviews, or other methods. At some point UNC SILS has not only a responsibility to guide students towards the collection and analyzation of that data, but the management of data as well. Data management is a growing area as we come to realize that the amount of data created will only increase exponentially (Poole, 2016).

2 Literature Review

In some scientific fields, data sharing has always been at the forefront of the work; some work is not really possible to do with the limited amount of data that one

group can collect. Weather data, for example, has been collected and shared for over 200 years (Sieber, 2015). Criminal justice data in the US is legion and very easy for anyone to access on the internet (see the National Archive of Criminal Justice Data, and the FBI's Crime Data Explorer for a few examples). Sharing data makes science better, as Sieber states,

> Data are the foundation of empirical research in all of the sciences. To understand and build on the work of others, researchers often need access to the data on which the work is based. Data sharing reinforces the norm of openness in scientific inquiry. It fosters verification, refutation or refinement of existing findings. It promotes new research, new ideas and alternative perspectives on any given scientific problem. It encourages more appropriate use of empirical data in policy making and evaluation. It fosters improvement of measurement and data collection methodology. And, it provides a powerful pedagogical tool for teaching students research design, analysis and interpretation of findings (Sieber, 2015).

The following literature review will start broadly with a discussion of the open data movement, and narrow to a discussion of data sharing and the role of the library in data sharing.

2.1 Open Data

The concept of data sharing is situated within the open data movement. The open data movement is conceptually about freeing data created by researchers from barriers imposed by publishers, for the most part - barriers of copyright and price (Molloy, 2011). The Open Knowledge Foundation defines open data as such: "Open data and content can be freely used, modified, and shared by anyone for any purpose" (The Open Definition, 2021).

The open data movement is fairly new. We start seeing more publications about the need for definitive open data policy in around 2007/2009. Uhlir and Schroder in 2007

attribute this to a shift to a more global science, more digital data, and an uncoupling of data collection from research.

> Researchers mostly collected and used their own data in their own research projects and had access to few external data sources. However, with the advent of digital technologies and networks, together with the growing scale and scope of research activities worldwide, the various parts of the research trajectory have been loosened into separate specialised activities (as, for example, data collection or technical support) that may be executed by different entities, in-house or outside the research institute (Uhlir & Schroder, 2007, p. 38).

They point out that this uncoupling allows data from multiple sources to be analyzed together to find unanticipated correlations and unpredicted results (Uhlir & Schroder, 2007). Murray-Rust makes a similar argument – that one benefit of open data is that researchers can "aggregate data from multiple papers" (2008) if the data is not copyrighted by the publisher in the form of supplemental information (Murray-Rust, 2008).

Other arguments for open data are that data need to be available to everyone to increase "transparency and reproducibility" (Molloy, 2011, p. 1) in research in order to make science better and more efficient for everyone. "Better science—in terms of transparency, reproducibility, increased efficiency, and ultimately a greater benefit to society—depends on open data" (Molloy, 2011, p. 4).

Open access and open data can open doors to greater diversity, equity and inclusion in science. Publications and data that are trapped behind a paywall only serve those that have the tools to access them (some type of institutional affiliation). Open data allows access to those outside the academy and allows distribution beyond just researchers at institutions that can afford the subscription (Ezema & Onyancha, 2016).

Data sharing can allow those that don't have access to the means to collect their own data, data to analyze. Open data allows these tools to be distributed more equitably. Researchers who can't access traditional publishing methods can still publish their data in open data repositories.

The Open Knowledge Foundation Working Group on Open Data in Science was formed in 2009 with a goal to advance the cause of open data in science with "guidelines, tools and applications" (Molloy, 2011, p. 1). They define open as both free from restrictions and as accessible and usable, and distinguish between "libre "free as in freedom" " (Molloy, 2011, p. 2) meaning free from copyright restrictions and "gratis "free as in beer" " (Molloy, 2011, p. 2) meaning free from cost. Molloy also discusses issues of access – data may be technically available but "requesting data from other researchers can be a torturous and sometimes fruitless process" (Molloy, 2011, p. 3). The Open Knowledge Foundation Working Group created a set of principles for the publishing of open data called the Panton Principles, with the goal to keep data in the public domain:

> 1. When publishing data, make an explicit and robust statement of your wishes.
> 2. Use a recognised copyright waiver or license that is appropriate for data.
> 3. If you want your data to be effectively used and added to by others, it should be open as defined by the Open Knowledge/Data Definition—in particular, non-commercial and other restrictive clauses should not be used.
> 4. Explicit dedication of data underlying published science into the public domain via PDDL (http://opendatacommons.org/licenses/pddl/1-0/) or CCZero (http://creativecommons.org/publicdomain/zero/1.0/) is strongly recommended and ensures compliance with both the Science Commons Protocol for Implementing Open Access Data and the Open Knowledge/Data Definition. (Molloy, 2011)

Open data today has advanced to the point that it is being used to research the current COVID-19 pandemic. Alamo et al enumerate over 20 different worldwide

institutions providing open data about the pandemic (including Google, the WHO, and the COVID-19 data hub); open source software communities doing work to ease access to the COVID-19 datasets; regional and international open datasets about COVD transmissions; and other non-COVID open data sets that are being used to support the study of COVID transmission including demographic, weather, and mobility datasets (Alamo, Reina, Mammarella, & Abella, 2020). Without open data researchers would not be able to follow this pandemic across all geographic locations as they have been.

As predicted by Murray-Rust, in a study of treatments for COVID-19, Zeng et al used deep learning and network-based techniques to study over 24 million research articles and create a knowledge graph to identify drugs that could be used to treat COVID-19 (Zeng, et al., 2020), work that wouldn't be possible at all without open data.

Open data all sounds great and very useful but all of the support in the world for open data matters not if people don't share their data.

2.2 Data Sharing

The benefits of data sharing are fairly well circulated – Tenopir et al provide a succinct list:

- re-analysis of data helps verify results data, which is a key part of the scientific process;
- different interpretations or approaches to existing data contribute to scientific progress –especially in an interdisciplinary setting;
- well-managed, long-term preservation helps retain data integrity;
- when data is available, (re-)collection of data is minimized; thus, use of resources is optimized;
- data availability provides safeguards against misconduct related to data fabrication and falsification;
- replication studies serve as training tools for new generations of researchers (Tenopir, et al., 2011).

Though the benefits are pretty well agreed on, this movement towards open data and data sharing is still unfolding and wholesale open sharing of all data is not something that is happening right now (Kim, 2013; Fecher, Friesike, & Hebing, What drives academic data sharing?, 2015). Fecher et al found that though a large majority of researchers studied felt that data sharing would benefit the research community and that other researchers should share data, only about 13% say they have shared data with the public (a higher percentage did state that they shared data ad hoc) (Fecher, Friesike, Hebing, & Linek, A reputation economy: how individual reward considerations trump systemic arguments for open access to data, 2017). In a 2018 study of PLoS journal articles (a publisher that requires a data availability statement and requires researchers to make all data available) and data availability statements, it was found that out of 47,661 articles studied 18.2% published their data in a named public repository (Federer, et al., 2018). Some data availability statements said the data was shared but didn't offer enough information to find the data, or held placeholder information that was never updated. The most common stated method of sharing was to say the data was in the paper or in the supplementary information, but of course, they were not referring to the entire dataset. (Federer, et al., 2018).

A primary reason that researchers seem to share data is in response to a requirement, either from their funder or from their publisher. Researchers generally agree that data sharing is good but don't actually do it for a number of reasons including intellectual property issues, and the time and work it takes to prepare data for sharing unless they are required to. Funder requirements are more of an inducement than journal

requirements. A study on data withholding found that 26% of survey respondents are "deterred from publishing articles if a journal requires the publication of data" (Fecher, Friesike, Hebing, & Linek, A reputation economy: how individual reward considerations trump systemic arguments for open access to data, 2017), insinuating that researchers that would rather not share data just choose a different journal. When it comes to funders, it is not as simple as just choosing a different one.

Funders wield a bit more power to require data sharing than journals do. In a 2015 study Higman & Pinfield reviewed 37 research data management policies at universities in the UK that were published in 2012 and 2013 and found that most were fairly vague, didn't define research data, and didn't discuss funding for research data management. They also found that most policies did mention funders and funder requirements, suggesting that the push to research data management and data sharing is coming from funder policies specifically (Higman & Pinfield, 2015). In another 2015 study of data sharing in the social sciences, Youngseek & Adler looked in part at how institutional pressures (including from funders and journals) influence data sharing and ultimately found that they didn't, though the survey was distributed in 2012 and the authors point out that many funders and journals had not yet began requiring data sharing (Youngseek & Adler, 2015).

Data sharing is fairly new and is something that researchers will have to do as requirements become more stringent and as the field matures.

2.3 Data Repositories and the library

When data is shared, it is most reliably shared in a public data repository. The entire field in general is in some transition with this. Those that agree that data sharing is good, and that do share their data, mostly share data on a personal website. In a study of derived astronomy data (different from primary data-which is widely shared) from articles published 1997-2008 researchers found that data is primarily shared via links in articles to websites where data can be found. The majority of links to personal websites of papers before 2004 are no longer available. Links that link to archival sites are more available (about 15-20% are broken). Researchers state "astronomers appreciate, but cannot reliably meet, the need to reference and include data materials in their published work in order to preserve its value" (Pepe, Goodman, Muench, Crosas, & Erdmann, 2014). The same study interviewed 12 researchers in astronomy and while they all said they would make their data available, only 2/3 actually do, and do so on personal websites (Pepe, Goodman, Muench, Crosas, & Erdmann, 2014). Personal websites are not the most reliable way to share data, as they generally don't outlast the person that runs them.

There are a number of issues with data not being archived in a data repository - data can be hard to find, requests for data can go unanswered, and data can be lost (Federer, et al., 2018). "Easy to use data repositories" are one suggestion that comes up in recommendations of how to make data sharing easier (Fecher, Friesike, & Hebing, What drives academic data sharing?, 2015). Data repositories not only store data, but can aid in ingest as well – making them a critical part of any data sharing team (Cragin, Palmer, Carlson, & Witt, 2010). They are also much better preservers of data, with

incentives and resources to migrate data to different formats as needed in future, as well as simply being more likely to be around (for instance, a library is more likely to be in existence in 50 years compared to a personal academic website).

Data repositories in universities often fall under the research data management umbrella, and are a part of the university library. In 2013 Kim studied the emerging role of academic libraries in data sharing and found that academic libraries have a strong role in data management in universities, including providing the repositories, providing the support in "developing and increasing data awareness within their institutions" (Kim, 2013, p. 501), and offering support to researchers regarding how to ready data for public sharing. As Kim states "it is imperative that libraries educate researchers on responsible data sharing and reuse practices through informal advice, consultation, class, and/or professional development" (Kim, 2013, p. 501). Libraries offer this support by providing guidance to researchers on the library website; providing workshops, consultations, and trainings; and expanding existing institutional repositories to include data (Kim, 2013).

In 2015 Cox et al sent a questionnaire to academic libraries in seven different countries to assess the current role of the library in research data management. They differentiated between research data management advisory services like trainings, planning assistance, and "web resource guides"; and technical services like data curation, metadata creation, and data repositories and found that most libraries surveyed didn't provide much in the way of advisory or technical services but that the area was expanding and incorporating with existing services. One of the identified "challenges" in growing research data management services was finding properly trained staff,

suggesting that there is a place here for library schools to ensure data management activities are taught (Cox, Kennan, Lyon, & Pinfield, 2017). Cox et al conclude that

> Although there were indications of significant leadership activity from the library community, there was also evidence of a less-developed service portfolio with much work still to be implemented. The scale and complexity of research data management support requirements mean that a wide range of services from advocacy to technical support, are needed at different stages of the research data lifecycle, and the skills and capabilities necessary are not consistently in place (Cox, Kennan, Lyon, & Pinfield, 2017, p. 2194).

A more recent 2020 study surveyed the libraries of research universities in the US to explore research data services (RDS) and found that the services considered most important from the library were "data archiving, data preservation and data documentation"; librarians also rated "assistance with funder mandates to be highly important" (Joo & Schmidt, 2021, p. discussion), but found services like data collection and analysis to be not as important (Joo & Schmidt, 2021). As the role of the library in RDS continues to mature, we can expect to see libraries providing those advisory services in data archiving and preservation and also hopefully building the technical services as well.

Research data management services can provide assistance to researchers in making sure their data is normalized and conforms to accepted data standards, including metadata creation, and the use of defined ontologies – both issues that are significant barriers in the re-use of shared data. Data needs to be normalized before it can be studied in aggregate. Standardization of metadata and ontologies ensures that different data sets are speaking the same language and allows reuse of data without the need to translate to other standards – losing information in the process (Thessen & Patterson, 2011).

Open data work is library work and a commitment to open data falls in line with many of ALA's Core Values of Librarianship, including access, democracy, intellectual freedom, the public good, service, social responsibility, and sustainability (Core Values of Librarianship, 2019). IFLA's (International Federation of Library Associations and Institutions) 2018 Global Vision Report found that librarians rated their highest value as "equal and free access to information and knowledge," and second highest as "commitment to dissemination of information and knowledge" (IFLA Global Vision Report: Library core values boost open science, 2018). Data sharing responds to each of these values, and ensuring that UNC SILS students graduate with some sense of how to share data is imperative.

2.4 Conclusion

Open data is an area where library science and information science truly converge. As academic libraries take charge of working with academic researchers to manage data, library and information scientists are at the forefront. As such, library and information science students should be open to open data and actively working to make sure data they create is openly available.

When we tie all of the above threads together, we come to the conclusions that open data is good for science, data sharing is an integral part of open data, data is best shared in public institutional repositories, and running institutional repositories is library work; these conclusions lead us to the larger conclusion that UNC SILS students need a firm grounding in data sharing as part of their education here at UNC SILS. One pretty easy, obvious, and low stakes way to cement that is to have a data sharing requirement as

part of the master's paper or project. This project will offer some best practices and guidelines in how to do that.

3 Methods

The purpose of this project is to create a set of documents to guide UNC SILS masters students in how they can publicly share data from their master's papers, and to propose to UNC SILS that they adopt a data sharing requirement in line with current requirements for funders and journals. I researched current data sharing policies of funders, publishers, and universities to create some policy recommendations, and then researched data sharing standards and processes to create the libguide for students, and then proposed the recommended policy changes to SILS.

3.1 Research

The first step in creating these documents and policy recommendations was to research current data sharing policy, and data sharing norms.

Extensive research on current data sharing policies of funders, journals, universities and other LIS programs was conducted in order to situate UNC SILS within the data sharing landscape and help to create a policy that works for the program. I read policies of major funders (public and private) and publishers, and tracked their data sharing requirements on a spreadsheet. I also searched to see if any other LIS master's programs have a data sharing requirement for the master's paper, and if so, what is required.

I filtered this information with a fairly simple "open coding" method (Strauss & Corbin, 1990). I read the policies, looked for similar themes, tracked those themes on a

spreadsheet, and reviewed the spreadsheet for items that most policies have in common (for example – requiring a data management plan). I then researched those common themes in the literature to see how they would fit with the needs of UNC SILS. I based my final recommendations to UNC SILS on this research.

To create the libguide on best practices and guidelines for data sharing, I researched existing data sharing standards, including metadata requirements and FAIR data standards; and looked into university institutional repositories to see what their requirements for shared data are, including the Carolina Digital Repository and Odum's UNC Dataverse.

3.2 Guidelines and best practices

Drawing from this research, I created a set of best practices for sharing UNC SILS master's paper/project research data. This 'best practices' document serves as a basic introduction on how to share data effectively. Following those best practices, I created a set of guidelines for SILS students to use to share their data. The 'guidelines' document serves as an explicit list of steps that students can use to prepare and to deposit data.

These guidelines and best practices are published as a libguide using a WordPress site at tarheels.live. The libguide can be accessed here: https://tarheels.live/uncsilsdatasharing/.

3.3 Evaluation

I enlisted the help of fellow SILS student, Carolyn Welker. In the style of a contextual interview, we worked together over Zoom to follow the guidelines to deposit

data into the CDR. I also reached out to Rebekah Kati at the Carolina Digital Repository for feedback on my data sharing guide. I used feedback from my colleague, myself, and the Carolina Digital Repository to further refine the guidelines.

Any feedback that I receive once I propose the guidelines and best practices to UNC SILS will also be incorporated. If the proposal is accepted and UNC SILS does accept a data sharing requirement for the master's paper I suspect that further evaluation would be in the hands of UNC SILS.

3.4 Propose to UNC SILS

To propose this idea to UNC SILS, I emailed this project to Rebecca Vargha - SILS Librarian, Aaron Brubaker - SILS Director of IT, and Brian Sturm - SILS Associate Dean for Academic Affairs with a brief cover letter that summarized the project and the recommendations.

3.5 work plan chart

	April 2021	May 2021	June 2021	July 2021
Research				
Write guide				
Create Webpage				
Deposit data				
Evaluate process				
edit guidelines				
propose to SILS				
complete and turn in				due July 29

4 Research/Results

4.1 Data sharing policy

Data sharing policy in the US is shaped by requirements of funders of research and by publishers of research. The university also has a large role in supporting data sharing through repositories and offering support to researchers, but the most cogent examples of data sharing requirements in the US come from funders and from publishers. Before such requirements data sharing existed but was more diffuse, though some fields were certainly more advanced in the data sharing sector than others. In the case of ecology data, the Long Term Ecological Research (LTER) Network issued guidelines for data sharing in 1990 that covered issues of availability, long term storage, and redistribution. In that same year, the LTER Network published a data catalog. By 2001 data sharing was the norm in the field (Porter, 2010), but this is not the case across the board. In a 2005 article studying what faculty members want and need from institutional repositories in order to share data researchers stated, “The phrase "if you build it, they will come" does not yet apply to IRs” (Foster & Gibbons, 2005), and in a nutshell, found that faculty want to share their work, including their data, but are very busy and data sharing work is mostly clerical work that cuts into research and writing time, and that they didn’t really understand the need to share data in a repository (Foster & Gibbons, 2005). Data sharing policy has driven researchers to begin to share their data.

4.1.1 Funders

The impetus for federal funding agencies to write policies that include data sharing requirements was the 2013 Memorandum from the US Office of Science and

Technology Policy (OSTP) that required that any "Federal agency with over $100 million in annual conduct of research and development expenditures to develop a plan to support increased public access to the results of research funded by the Federal Government" (Holdren, 2013, p. 2); and "the results of unclassified research that are published in peer-reviewed publications directly arising from Federal funding should be stored for long-term preservation and publicly accessible to search, retrieve, and analyze" (Holdren, 2013, p. 3). Before the 2013 memo, a 1999 memo stated that research data that came from federally funded projects be made public under the Freedom of Information Act (OMB, 1999). More recently, dealing with data created by government agencies, the Foundations for Evidence Based Policy Making Act of 2018 was passed, intending to "improve Federal data management" (Foundations for evidence based policy making act of 2018, 115th Congress Public Law 435, 2019) and increase access to open government data. Title II of the above law is called the Open Government Data Act and calls to make government data open data by default.

Due to the 2013 memorandum, all federal agencies now have some type of data sharing plan, referred to generally as a public access plan or data management policy, that details how the research papers and data that come from federally funded research or research done by their employees will be shared. While the plans all differ somewhat, most requirements are similar. Most require a data management plan that outlines the type of data to be created and how it will be generated, shared, and preserved; and if data sharing is required that it be shared in a public repository. All plans place at least some limits on what data must be shared publicly.

The National Science Foundation (NSF) and the National Institute of Health (NIH) both require data management plans – their requirements are similar to requirements of other federal agencies. The National Science Foundation requires a data management plan (DMP) of no more than 2 pages (Dissemination and Sharing of Research Results - NSF Data Management Plan Requirements, 2021) that covers the types of data and materials the project will produce, intended standards for data and metadata, how the data will be shared and how it can be accessed - to include issues of intellectual property, privacy, security issues, policies for re-use and "production of derivatives," and data archiving plans. After the research is completed, data is required to be deposited into a public repository as outlined in the DMP unless the data is not required to be shared. Some allowances are made for proprietary data, confidential and personally identifiable information, classified information, and information that may pose a national security issue. The data management plan should address any of these issues (NSF 15-052, Today's data, tomorrow's discoveries, 2015). Similarly, in the new Final NIH Policy for Data Management and Sharing that will be effective starting in 2023, the National Institutes for Health states specific requirements for the data management plans. The plan should be two page or less and should address the type and amount of expected data, what data will be shared and why, the metadata and other information that will be shared along with the data, a list of other tools that may be needed to access the data (like software) and details on how to access it, a timeline for depositing the data into a repository and information on which repository will be used, a discussion of security and privacy issues, and how and who will monitor the data plan. The NIH also encourages the use of a data repository for sharing data and provides some guidance in how to choose a

data repository (NOT-OD-21-014; , 2020). These changes bring the NIH plan in line with the NSF plan by stating specific requirements for the data management plan and requiring it for all projects and not just those over a certain financial threshold.

Most agencies require data sharing as outlined within the scope of the data management plan, but not all. The Department of Energy states that they will assess each plan and "take into account the relative values of long-term preservation and access and the associated cost and administrative burden" before deciding if the data must be shared (U. S. Department of Energy Public Access Plan, 2014).

Each agency does impose some limits on data sharing and don't require all grantees to share all data across the board. Some limits are broader than others but most don't require sharing of data that is confidential, classified, or intellectual property. The Department of Veterans Affairs goes a step further to state that all data must be shared, but some data may be shared privately – differentiating between "open public access" and "controlled public access" (Department of Veterans Affairs Policy and Implementation Plan for Public Access to Scientific Publications and Digital Data from Research Funded by the Department of Veterans Affairs, 2015, p. 13). The Assistant Secretary for Preparedness and Response (ASPR) policy is pretty clear and representative of other plans with these limits:

> Digital scientific data that are not in scope for this plan include:
> • Personally identifiable data
> • Proprietary trade data
> • Data related to protecting critical infrastructure
> • Other data whose release is limited by law, regulation, security requirements, or policy (Response, 2015).

Federal public access plans either suggest a repository, provide a data repository, state requirements for a repository, and/or provide suggestions and offer assistance in identifying a good repository. The Administration for Community Living lists a preferred repository (the ICPSR at University of Michigan), but researchers can choose another public repository if desired (Administration for Community Living public access plan, 2017). The US Department of Agriculture has a repository called the AG Data Commons - https://data.nal.usda.gov/, though again it doesn't look like researchers are required to use it, they may choose another public repository (Implementation Plan to Increase Public Access to Results of USDA-funded Scientific Research, 2014). For more information on the public access policies of federal funders, see appendix A.

Some private funders require sharing of data (American Heart Association, Gates Foundation, Wellcome), but generally private foundations express a commitment to open access, with policies that tend to focus more on open publication of the final results of grants, and intellectual property licensing, and less on data and data sharing. Some private funders state that products of funded research should be shared and licensed under a Creative Commons license (Gates Foundation, Ford Foundation, Hewlett Foundation, Sloan Foundation), and that publications resulting from grant funds should be accessible to the public (HHMI, Moore Foundation, MacArthur Foundation, Wellcome, and the American Heart Association) but stop short of saying all data should be shared. This is potentially due to a number of factors – namely the lack of a clear-cut overriding requirement such as the 2013 OSTP memo that all public funders cite, and the fact that work funded by private funders is so much more broad – ranging from empirical research

to digital projects, both in the US and internationally. For more information in private funders public access plans see appendix B.

4.1.2 Publishers

All major publishers have some type of data policy and provide some level of support for data sharing, but again, most stop short of requiring public data sharing, though there are exceptions (Plos, Emerald Open Research). Some publishers specifically state a requirement that data must be shared with other researchers who request the data (AAAS Science Journals, AGU, JMIR). Other publishers offer a number of options that each journal can choose from – ranging from encouraged to share, to required to share plus peer review of data (SpringerNature, Sage, Taylor and Francis, Wiley). One good example is the SpringerNature policy (which itself is available under the Creative Commons attribution license). SpringerNature has 4 policy levels:

> Type 1: Data sharing and data citation is encouraged
> Type 2: Data sharing and evidence of data sharing encouraged
> Type 3: Data sharing encouraged and statements of data availability required
> Type 4: Data sharing, evidence of data sharing and peer review of data required (SpringerNature Research Data Policy Types).

In our research on funders we found that many required a data management plan - for publishers we see that many require or suggest a data availability statement which is a statement in the final article that tells the reader what data is shared and where that data can be found. JMIR offers 5 potential forms for the data availability statement:

> Data Availability statements can take one of the following forms (or a combination of more than one if required for multiple data sets):
>
> - "The data sets generated during and/or analyzed during the current study are available in JMIR Data | the [NAME] repository, [PERSISTENT WEB LINK TO DATA SETS]"

- "The data sets generated during and/or analyzed during the current study are not publicly available due [REASON WHY DATA ARE NOT PUBLIC] but are available from the corresponding author on reasonable request."
- "The data sets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request."
- "Data sharing not applicable to this article as no data sets were generated or analyzed during the current study."
- "All data generated or analyzed during this study are included in this published article [and its supplementary information files]." (JMIR, 2021).

The publishers that don't explicitly require data sharing do state strong encouragement for it and offer resources to assist in sharing, including manuals, and repositories or advice on finding repositories. For example, the APA provides a repository (https://osf.io/meetings/apa/) and offers resources on their website (https://www.apa.org/pubs/journals/resources/data-sharing). The Emerald Open Research Data Guidelines include step by step instructions and guidance on how and where to share data (https://emeraldopenresearch.com/for-authors/data-guidelines) (Data Guidelines, 2021). For more information in public access plans of publishers, see appendix C.

4.1.3 Universities

In researching data sharing policies of the top 10 library science schools (based on the US News and World Reports list (https://www.usnews.com/best-graduate-schools/top-library-information-science-programs/library-information-science-rankings), it was found that university data policies don't really cover research data sharing specifically. University data policies are geared more towards security of university held data like student data. Some have research data policies that cover ownership and

management of research data created within the university. Most universities researched do have open access policies that apply to journal articles and conference proceedings, but not data, of faculty and staff that require faculty/staff to grant the University a license to the final materials so that they can be stored in the university repository (these policies generally state that the author retains copyright). For example, the UNC Open Access Policy calls for all faculty members to grant the university a license to all of their "scholarly articles" in order to make them freely available in an open access repository (UNC Open Access Policy, 2016).

In researching master's paper/thesis requirements of other library and information science schools, no schools were found to have any kind of data sharing/data management requirement for the master's paper/thesis. Though all schools researched did have an institutional repository for the deposit of the paper/thesis itself, and most did have data repositories, a quick look through the data held suggested that those repositories did not appear to hold a large amount of masters paper/thesis data.

Though universities themselves don't specifically require sharing of data, universities do have a place at the data sharing table, but not in policy making – in offering support services like repositories, research data services departments, and consulting services. For example, the Odum Institute at UNC offers assistance with research data throughout the data lifecycle – from the proposal stage, through data collection and analysis, to data archiving. The University does play a key role in open data and data sharing by making the sharing of data possible and equitable for all affiliated with the university – faculty, staff, and students. For more information on data sharing and LIS master's programs, see appendix D.

4.1.4 Discussion/Application

In considering data sharing policy for a university master's program, and taking all other policies into account, it would be best to stay broad (in the example of the private funders and publishers), and require that students consider data, but not necessarily share data. The federal funders have more stringent guidelines for data sharing with the public and rightfully so, the public should have access to data that is funded by public money. Private funders and journals have more leeway in this regard and so should UNC SILS students. An understanding of the data created and how it should be managed is useful to students and can be accomplished by requiring students to provide a data management plan with the proposal. Readers of the papers will benefit from a data availability statement that explains where the data is located. Actual sharing of data should not be required.

At UNC SILS, students deposit their master's papers/projects in the Carolina Digital Repository and they already have the option to deposit data, but most don't. Providing the students with the tools and guidelines to deposit data should increase the amount of student data deposited in the CDR.

I recommend that SILS adopt a policy that requires a data management plan created with a standardized DMP tool as part of the proposal, and a data availability statement as part of the final paper/project. Neither of these recommendations are a particularly heavy lift for students, but will be useful to readers of the papers and will keep SILS and SILS students current with the industry trends towards open access and data sharing.

The following sections cover the research on the guidelines for creating the data management plan and data availability statement. Students won't be required to actually share data, but if they do, the guidelines below will provide an easy to follow framework and suggested resources.

4.2 How to share master's paper data

In addition to policy recommendations, the second result of this master's project is a libguide for master's students with information needed to respond to the above policy recommendation to create a data management plan and data availability statement, and to share their data if desired. The libguide will include information on best practices in data sharing, and will outline the steps to share data at UNC, including how to create a data management plan and data availability statement. The website will focus on data sharing through the Carolina Digital Repository (CDR), but will include links to other repositories as well.

This section will cover the research that goes into that libguide. Best practices will cover data sharing norms like what accompanying information and metadata should be included with you data deposit, formats data should be deposited in, and reuse rights; and the step by step guide will cover how to select a data repository, how to write a data management plan, how to deposit data into a repository, and how to write data availability statement and where to include it.

4.2.1 Why, or why not share your data

Publishing data sets in open repositories allows access to those beyond the academy and expands the reach of scholarship. Open data and open access in general can

contribute to a more diverse knowledge base that can be accessed and studied by more people - open data can contribute to more diversity of thought (Kittinger, 2020).

The other side of that is that data can be used to harm. The data can be misunderstood to make conclusions about populations that are incorrect, and data can leave out vulnerable populations entirely (Qureshi, 2020). There are a number of other reasons that it would not be prudent to share data. If the data contains personally identifiable information of any kind, it should be anonymized completely before sharing, or not be shared at all (the CDR will not accept any data with personally identifiable information). If the data was collected from humans (either via survey or interview for example) it would be prudent to ask them if they are ok with the data being anonymized and shared before you share it. If your data contains copyrighted material or material that is the intellectual property of someone else – you might not be able to share it (depending on the restrictions of the copyright/license).

4.2.2 Choosing a repository

If you decide to share data, choosing a repository is a reasonable first step as this information goes into the data management plan and data availability statement and will guide some other choices, like metadata and other documents to prepare.

A repository is a collection of articles or data that pertain to a specific thing, some are institutional and encompass the works of the institution – in the case of a university, the work of faculty and students, other repositories are subject specific and encompass the work of a particular subject.

UNC students have access to two different institutional repositories, the Carolina Digital Repository (CDR) and Odum Institute's UNC Dataverse. Both are available to UNC students to deposit but have some key differences. The UNC Dataverse allows users to work with qualitative data online, while the CDR allows for embedding audio/visual files. The CDR requires an ONYEN so students won't be able to change their deposit themselves after graduation. Other considerations are outlined here: https://guides.lib.unc.edu/researchdatatoolkit/choosing-a-repository

There are a number of reasons to choose a subject matter repository over an institutional repository, but not many really apply here. For example – genetic sequences must be deposited into GenBank, and protein structures must be deposited into Protein Data Bank (Emerald Open Research How to Publish, 2021). Re3data.org offers a directory of subject specific repositories for students interested in exploring that option https://www.re3data.org/search. This guide will focus on the CDR because student master's papers/projects are already deposited in the CDR and the data can be linked to the paper.

4.2.3 Data Management Plan

A data management plan (DMP) can be a useful tool in making decisions about the type of data and metadata to collect, and mapping how the data will be analyzed and stored. The data management plan goes into detail about the types of data to be collected, and specifics on how the data will be collected and maintained and then archived. The data management plan can also include information on why the data won't be shared – issues of information confidentiality, copyright, and security clearance.

The use of a standard tool for creating a DMP has a number of benefits. Mainly, it is just easier for students, the DMP tool lays out exactly what to include and then creates a plan in a standard format. Also, the DMP tool creates plans that all follow the same format, and so are interoperable and can be studied themselves; the plans can become data. One interesting line of inquiry created by the requirement to add a data management plan to the master's paper/project proposal is an empirical study of the plans themselves. To this end, I suggest that data management plans be created using the standardized DMP Tool from the Digital Curation Center (see (Bishoff & Johnson, 2015) (Rolando, et al., 2015) for studies that compare plans). UNC Chapel Hill is a participating institution, and as such students can sign in via the shibboleth. The DMP Tool is simple to use and asks for some basic information:

> What data will you collect or create?
> How will the data be collected or created?
> What documentation and metadata will accompany the data?
> How will you manage any ethical issues?
> How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?
> How will the data be stored and backed up during the research?
> How will you manage access and security?
> Which data are of long-term value and should be retained, shared, and/or preserved? What is the long-term preservation plan for the dataset?
> How will you share the data?
> Are any restrictions on data sharing required?
> Who will be responsible for data management?
> What resources will you require to deliver your plan? (DMP Tool, 2021).

The DMP tool offers guidance on how to answer each question. Students enter their answers into an online form, and the DMP finalizes the plan, and creates a downloadable version.

4.2.4 General Data sharing norms

When considering sharing data there are some standard practices to consider so that the data can be useful to as many people as possible. An oft referred to general set of guiding principles are the FAIR data principles (Wilkinson, 2016), created to help researchers ensure that their data is accessible. We will discuss data sharing norms within the scope of these FAIR principles that are:

> Data should be **Findable**
> Data should be **Accessible**
> Data should be **Interoperable**
> Data should be **Re-usable**
> (Force11, 2021).

The distinctions between these principles can blur at times, but generally findable refers to ease of locating the data, accessible refers to ease of reading the data, interoperable refers to ease of working with the data, and re-usable refers to issues that impact the ability of others to reuse the data including issues of copyright, provenance, and standards (Wilkinson, 2016). Metadata is key to fulfilling all of these elements.

4.2.4.1 Metadata

"Metadata is a map" (Pomerantz, 2015). Metadata is important in that it creates descriptive access points that allow other researchers to locate the data, and administrative information that allows others to understand how to interpret the data. It is also useful to provide some description of how and why the data was created (Pomerantz, 2015). This metadata and descriptive information should be included with the repository deposit.

In order for the metadata to be understood by humans and machines, it is useful to use an already defined metadata standard. Dublin Core is a popular and basic standard that was created to fulfil any metadata need. The original Dublin Core metadata element set includes 15 elements for descriptive metadata – these elements and definitions can be found at https://www.dublincore.org/specifications/dublin-core/dces/ and range from information about the contributor to description of the resource to rights information. The basic Dublin Core element set can be expanded if desired using the expanded DCMI metadata terms here https://www.dublincore.org/specifications/dublin-core/dcmi-terms/. The Library of Congress manages two other fairly basic metadata schemas/elements – MODS (https://www.loc.gov/standards/mods/mods-outline-3-7.html) intended mainly for library applications and METS (https://www.loc.gov/standards/mets/METSOverview.v2.html). A significant number of more specialized standards can be found here: https://fairsharing.org/standards/. Many repositories have already defined the metadata that is required for the repository deposit, other metadata can be included in a text file with the deposit if desired.

When depositing into the CDR the metadata fields are already defined, and the CDR creates a machine-readable version of said metadata for indexing. The CDR also asks for a document that explains how the data can be accessed and reused, this document can also include additional metadata and an ontology if needed. The CDR requires the following metadata fields: title, creator #1, date of publication, abstract, methods, kind of data, and resource type (with a list of terms to choose from); and gives the option to add some optional fields including another creator, and more information about the creator, and contributor (and info about the contributor), keyword, license, subject, language,

location, related resource url, funder, last modified date, project director (and information about the project director), researcher (and information about the researcher), and sponsor (CDR Add a new dataset, 2021).

4.2.4.2 Findable

In order for shared data to be widely useful for research, researchers must be able to locate the data in a search, or from a citation. In order to fulfill this principle, data should be shared in a public repository with a persistent identifier (like a DOI number), and metadata that describes the data well enough for it to be located by researchers (Wilkinson, 2016).

DOI stands for digital object identifier and is intended to be a persistent identifier that points to a digital resource. The DOI is part of the citation of the digital object and allows the object to be findable over time (doi.org, 2015). The DOI makes the data easier to cite, and easier to find with a citation. The Carolina Digital Repository will provide a DOI number if the data is made public. It will be created automatically overnight and will be available in the DOI field of the deposit the next day.

For researchers to locate your data in a search, is it essential that the keyword metadata is robust enough to allow for finding in a web search. This can be accomplished through the use of an established thesaurus. In order to do this it is important to use keywords that come from a thesaurus or ontological schema appropriate for the field. The CDR uses FAST subject headings from the OCLC (https://fast.oclc.org/searchfast/) for subject keyword metadata. An extensive list of ontologies can be found at this libguide from Indiana University: https://info.sice.indiana.edu/~dingying/Teaching/S604/OntologyList.html.

4.2.4.3 Accessible

In order to be accessible data should be free of cost and copyright restrictions, and should be readable and downloadable through well-defined protocols (Force11, 2021). If the data is stored in a public repository with an open access license it will be both free of cost and (relatively) free of copyright restrictions. Administrative metadata information needed to download and use the data should be defined in the materials. The CDR provides an optional metadata field for copyright license information.

Creative Commons copyright licenses are well known open access licenses and are easily adaptable to be used with datasets. The Creative Commons licenses are:

> CC BY**:** This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use.
>
> CC BY-SA: This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use. If you remix, adapt, or build upon the material, you must license the modified material under identical terms.
>
> CC BY-NC: This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format for noncommercial purposes only, and only so long as attribution is given to the creator.
>
> CC BY-NC-SA: This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format for noncommercial purposes only, and only so long as attribution is given to the creator. If you remix, adapt, or build upon the material, you must license the modified material under identical terms.
>
> CC BY-ND: This license allows reusers to copy and distribute the material in any medium or format in unadapted form only, and only so long as attribution is given to the creator. The license allows for commercial use.
>
> CC BY-NC-ND: This license allows reusers to copy and distribute the material in any medium or format in unadapted form only, for noncommercial purposes only, and only so long as attribution is given to the creator.
> (About CC Licenses, 2021)

The Open Knowledge Foundation also offers some open licenses for data/databases:

> Open Data Commons Open Database License (ODbL) — "Attribution Share-Alike for data/databases"
>
> Open Data Commons Attribution License — "Attribution for data/databases"
>
> Open Data Commons Public Domain Dedication and License (PDDL) — "Public Domain for data/databases"
> (Open Data Commons Licenses, 2021)

The CDR offers licenses that track pretty well with the Creative Commons licenses:

> Attribution 3.0 United States
> Attribution-ShareAlike 3.0 United States
> Attribution-NonCommercial 3.0 United States
> Attribution-NoDerivs 3.0 United States
> Attribution-NonCommercial-NoDerivs 3.0 United States
> Attribution-NonCommercial-ShareAlike 3.0 United States
> Public Domain Mark 1.0
> CC0 1.0 Universal
> All rights reserved (CDR Licenses in the CDR, 2021)

In order to be downloadable through "well-established protocols" (Force11, 2021) it is helpful for the data to be in a format that does not depend on a specific software to open, and if it does, that software is open access so that all may use it free of charge. Stanford Libraries offers some guidelines for accessible file formats and include this table of preferred formats:

> Containers: TAR, GZIP, ZIP
> Databases: XML, CSV
> Geospatial: SHP, DBF, GeoTIFF, NetCDF
> Moving images: MOV, MPEG, AVI, MXF
> Sounds: WAVE, AIFF, MP3, MXF
> Statistics: ASCII, DTA, POR, SAS, SAV

Still images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP
Tabular data: CSV
Text: XML, PDF/A, HTML, ASCII, UTF-8
Web archive: WARC
(Best practices for file formats, 2021).

4.2.4.4 Interoperable

Data is considered interoperable if it is machine-readable, and uses commonly adopted terminology from an already existing ontology – if no such ontology exists, the researcher should include definitions for terms (Force11, 2021). Machine-readable data is "data that can be automatically read and processed by a computer" (Machine Readable, 2021). Raw data itself is hard for humans to look at and parse – the data must be in a format that can be processed by computers, analyzed, and displayed in a human readable form. CSV is the most widely used machine-readable format, but other formats can be machine readable as well (A Primer on Machine Readability for Online Documents and Data, 2012). The CDR provides machine readable metadata for each data deposit.

Information about the ontology used, and/or term definitions can be included in the codebook/readme that the CDR requires.

4.2.4.5 Re-Usable

Data is considered to be reusable by other researchers by FAIR standards if it conforms to the first three principles, if the metadata is robust enough to be able to be linked with other data sources, and if enough citation information is provided to be able to locate the data.

4.2.5 Deposit in Repository

This section will focus on the CDR, though some steps are useful regardless of repository, especially the preparation of the data.

The steps to deposit data into the CDR are well outlined on their website. The first step is to gather the data and prepare it per the data sharing norms above and ready it for deposit, including preparation of the readme/codebook document that the CDR requires to go along with the data. Once that is ready, students can follow the steps on the website to deposit a dataset. Students will add files, fill in the metadata, select the visibility level (public or restricted), and agree to the deposit agreement. https://blogs.lib.unc.edu/cdr/how-to-contribute-material-to-the-cdr/how-to-deposit-a-dataset-in-the-cdr/

The instructions mention depositing multiple files as child works. The student master's paper/project dataset would technically be a child work attached to the master's paper, but due to the workflows currently in place for the deposit of masters' papers (students deposit the papers, but they are held until the SILS library approves the submission – so students don't actually have ownership of the paper), datasets will need to be deposited as not child works and someone at the CDR will have to attach the dataset and paper. Students can email the general CDR email address (cdr@unc.edu) to have the dataset attached to the paper in the repository. Alternatively, a change in SILS procedure could be made to allow students to "own" their paper in the repository in order to attach the dataset as a child work.

If students are depositing into the UNC Dataverse they should visit this URL https://dataverse.unc.edu/, and click the ‘add data’ button that is on the right, just above the beginning of the dataset list. They will be prompted to create an account. After an account is created, they would click ‘add data’ and then ‘New Dataset’. More information on adding a dataset to the Dataverse is available on their user guide here: https://guides.dataverse.org/en/4.9.4/user/.

If students choose a subject matter repository, they can figure that out on their own, the basics should be similar though in regards to preparation of data and metadata and inclusion of a readme/codebook document.

4.2.6 Data Availability Statement

The data availability statement, as discussed earlier, is a short statement to readers of the paper that tells them how and where to locate the data if applicable, and should be included in the final version of the paper. Some examples were noted earlier in this paper. Springer Nature also offers some examples and templates in their website, as the Springer Nature policy is itself licensed under the Creative Commons attribution license – students can copy these data availability statements, edited as needed, as long as they cite Springer Nature in the paper.

> The datasets generated during and/or analysed during the current study are available in the [NAME] repository, [PERSISTENT WEB LINK TO DATASETS].
>
> The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.
>
> All data generated or analysed during this study are included in this published article (and its supplementary information files).

> The datasets generated during and/or analysed during the current study are not publicly available due to [REASON(S) WHY DATA ARE NOT PUBLIC] but are available from the corresponding author on reasonable request.
>
> Data sharing not applicable to this article as no datasets were generated or analysed during the current study.
>
> The data that support the findings of this study are available from [THIRD PARTY NAME] but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of [THIRD PARTY NAME]. (SpringerNature Research Data Policy Types).

The data availability statement would go after the conclusion in the final paper. If students are depositing data, it should be deposited before the final paper is turned in so that the link to the data can be included in the paper.

5 Guidelines and Best Practices

This libguide was created as a Wordpress site on tarheels.live and can be easily transferred to any platform. My hope is that this libguide can become a part of the master's paper/project documentation to help future students share data.

LINK: https://tarheels.live/uncsilsdatasharing/

6 Evaluation

The evaluation of this libguide was completed in two steps. First, a Systems Analysis style contextual interview was conducted with Spring 2021 UNC SILS graduate, Carolyn Welker. Once the libguide was edited based on that interview, the libguide was sent to UNC Institutional Repository Librarian, Rebekah Kati, for review.

Carolyn reviewed the "Best Practices" on her own to determine if she should share her data, and if so, to prepare her data to share. We then conducted a Zoom interview call together while she went through the steps in "How to share your data – step by step" to determine the utility of the process. Carolyn ultimately decided not to share her data due to privacy concerns for her interview subjects. Her feedback on the website was invaluable and her comments are summarized in the table below along with the changes that they precipitated.

SECTION	ISSUE	CHANGE MADE
Best Practices, Sharing data – Why and why not	Didn't include Carolyn's reasons for not sharing data	added more information about informing research subjects/interviewees that data will be shared
How to share your data – step by step, prepare for deposit	codebook link is broken	fixed it
	link to example of completed codebook may be useful	found link to completed codebook in CDR and added it
	definition of CDR metadata field may be useful	added further definition to metadata fields
How to share your data – step by step, Deposit to CDR	Child work part is confusing	explained that better
How to share your data – step by step, Attach data to paper in CDR	this whole part is confusing	got a better understanding of how that will work and revised that section
How to share your data – step by step, Data availability statement	say why and where in the paper this goes	revised to explain better

Rebekah Kati reviewed the website to ensure that the steps for deposit were correct and offered some guidance on how students could ensure that their data gets

connected to their paper in the repository. Using her feedback, I was able to clarify the deposit process a little more.

7 Proposal to UNC-SILS

After reading all of the above I hope that you will join me in the conclusion that SILS should include some level of data sharing with the master's paper/project requirements. I propose that SILS add a requirement for a data management plan to be included in the master's paper/project proposal, and a data availability statement to be included in the paper itself, actual data sharing would remain optional. These requirements could be easily added to existing master's paper/project requirements.

The requirement to include a data management plan can be included in the INLS 781, Proposal Development class materials. Currently the proposal consists of three sections, an introduction, a literature review, and a methods section. The data management plan could become the fourth required section of the proposal.

The requirement to include a data availability statement can be added to the "writing the text" requirements here after "d) Margins": https://sils.unc.edu/student-services/masters-students/masters-paper/guidelines#writing-the-text, the data availability statement would go at the end of the paper, after the conclusion. Suggested text:

> e) Data availability statement: after the conclusion and before the works cited include the data availability statement – a brief statement of a sentence or two to state what data was collected and where that data can be found. If no data was collected, or the data is not shared publicly, that information should be reflected in the data availability statement.

The website of best practices and steps for data sharing could be added to the SILS website with the masters paper requirements as well, in order to guide students who wish to share their data.

8 Conclusion

The benefits of data sharing have been well expounded in the literature review, in summary: sharing of data allows data to be reviewed to verify findings and to create new avenues of inquiry, data can be aggregated to explore different areas, data sharing minimizes the need to continuously collect the same data over and over again, and shared data is preserved in ways that unshared data is not. The data sharing movement is growing but is not mature, and many researchers are reluctant to share data, some for valid reasons (copyright, personally identifiable information), and some because of a lack of understanding about the process of how to share data. There is a place in this growing field for library and information science students to make a mark. That is why this proposal that UNC SILS students have some data sharing requirements for the master's paper/project is so important. It is an opportunity for them to gain first hand experience with data sharing practices.

Opening up access to the data that students collect for papers and projects can be an important step for UNC SILS to stay at the forefront of data trends that are becoming solidified into practice. The main benefits are the benefit to students in learning about the process of data sharing and to future students and researchers in the new lines of inquiry that data sharing will allow. Future students can browse through available data when planning master's papers/projects and use either single or aggregate datasets to form new

and interesting research questions. Why should we continue to create data and then keep it closely held once the paper is done when there is so much more that can be learned from that data?

This project seeks to increase data sharing among UNC SILS students by requiring them to take some steps to think about the data they will create and how that data will be managed, and to communicate that to readers of the paper. This project also gives students a helpful resource to follow to share their data if desired.

9 Data Availability Statement

The datasets in the form of csv spreadsheets of policy information from funders, publishers, and universities generated during and/or analyzed during the current study are available in the Carolina Digital Repository at https://cdr.lib.unc.edu/concern/data_sets/gm80j4446. DOI: https://doi.org/10.17615/765f-wa04

(SpringerNature Research Data Policy Types)

10 Works Cited

A Primer on Machine Readability for Online Documents and Data. (2012). Retrieved from Data.gov: https://www.data.gov/developers/blog/primer-machine-readability-online-documents-and-data

About CC Licenses. (2021). Retrieved from CreativeCommons.org: https://creativecommons.org/about/cclicenses/

Administration for Community Living public access plan. (2017). Retrieved from acl.gov: https://acl.gov/sites/default/files/about-acl/2017-12/ACLPublicAcccessPlan.pdf

Alamo, T., Reina, D., Mammarella, M., & Abella, A. (2020). Open data resources for fighting COVID-19. *arXiv:2004.06111v3*. Retrieved from https://arxiv-org.libproxy.lib.unc.edu/abs/2004.06111v3

Best practices for file formats. (2021). Retrieved from Stanford Libraries: https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats

Bishoff, C., & Johnson, L. (2015). Approaches to data sharing: an analysis of NSF data management plans from a large research university. *Journal of Librarianship and Scholarly Communication, 3(2)*, eP1231. doi:http://dx.doi.org/10.7710/2162-3309.1231

Bush, V. (1945). As we may think. *The Atlantic*. Retrieved from https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/

CDR Add a new dataset. (2021). Retrieved from Carolina Digital Repository: https://cdr.lib.unc.edu/concern/data_sets/new?locale=en

CDR Licenses in the CDR. (2021). Retrieved from Carolina Digital Repository: https://blogs.lib.unc.edu/cdr/licenses-in-the-cdr/

Core Values of Librarianship. (2019, January). Retrieved from American Library Associations: http://www.ala.org/advocacy/intfreedom/corevalues

Cox, A., Kennan, M., Lyon, L., & Pinfield, S. (2017). Developments in research data management in academic libraries: Towards an understanding of research data service maturity. *Journal for the Association for Information Science and Technology*, 2182-2199. doi:https://doi.org/10.1002/asi.23781

Cragin, M., Palmer, C., Carlson, J., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences, 368(1926)*, 4023-4038. Retrieved from http://www.jstor.org/stable/25704697

Data Guidelines. (2021). Retrieved from Emerald Open Research, How to publish: https://emeraldopenresearch.com/for-authors/data-guidelines

Department of Veterans Affairs Policy and Implementation Plan for Public Access to Scientific Publications and Digital Data from Fresearch Funded by the Department of Veterans Affairs. (2015, July 23). Retrieved from va.gov: https://www.va.gov/ORO/Docs/Guidance/VA_RSCH_DATA_ACCESS_PLAN_07_23_2015.pdf

Dissemination and Sharing of Research Results - NSF Data Management Plan Requirements. (2021, March 2018). Retrieved from National Science Foundation: https://www.nsf.gov/bfa/dias/policy/dmp.jsp

DMP Tool. (2021). Retrieved from dmptool.org: https://dmptool.org/

doi.org. (2015, October 17). *DOI Handbook*. doi:10.1000/182

Emerald Open Research How to Publish. (2021). Retrieved from Emerald Open Research: https://emeraldopenresearch.com/for-authors/data-guidelines#hosting

Ezema, I., & Onyancha, O. (2016). Status of Africa in the global open access directories: Implications for global visibility of African scholarly research. *Fourth CODESRIA Conference on Electronic Publishing.* Dakar, Senegal: CODSRIA.

Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing? *PLoS ONE 10(2) e0118053*. doi:https://doi.org/10.1371/journal.pone.0118053

Fecher, B., Friesike, S., Hebing, M., & Linek, S. (2017). A reputation economy: how individual reward considerations trump systemic arguments for open access to data. *Palgrave Communications 3*, 17051.

Federer, L., Belter, C., Joubert, D., Livinski, A., Lu, Y., Snyders, L., & Thompson, H. (2018). Data sharing in PLOS ONE: An analysis of data availability statements. *PLoS ONE 13(5)*, e0194768. doi:https://doi.org/10.1371/journal.pone.0194768

Force11. (2021). *Guiding principles for findable, accessible, interoperable and re-usable data publishing version b1.0.* Retrieved from Force11: https://www.force11.org/fairprinciples

Foster, N., & Gibbons, S. (2005). Understanding faculty to improve content recruitment for institutional reposiories. *D-Lib Magazine, 11(1)*. Retrieved from http://www.dlib.org/dlib/january05/foster/01foster.html

Foundations for evidence based policy making act of 2018, 115th Congress Public Law 435. (2019, 01 14). Retrieved from https://www.congress.gov/bill/115th-congress/house-bill/4174

Higman, R., & Pinfield, S. (2015). Research data management and openness: The role of data sharing in developing institutional policies and practices. *Program, 49 (4)*, 364-381.

Holdren, J. (2013, 02 22). MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES; Increasing Access to the Results of Federally Funded Scientific Research. Office of Science and Technology Policy. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

IFLA Global Vision Report: Library core values boost open science. (2018, September 18). Retrieved from ZBW Media Talk: https://www.zbw-mediatalk.eu/2018/09/ifla-global-vision-report-library-core-values-boost-open-science/

Implementation Plan to Increase Public Access to Results of USDA-funded Scientific Research. (2014). Retrieved from usda.gov: https://www.usda.gov/sites/default/files/documents/USDA-Public-Access-Implementation-Plan.pdf

JMIR. (2021). *What is your data sharing policy?* Retrieved from JMIR Publications Knowledge Base and Help Center: https://support.jmir.org/hc/en-us/articles/360030832631-What-is-your-data-sharing-policy-

Joo, S., & Schmidt. (2021). Research data services from the perspective of academic librariens. *Digital Libary Perspectives, Vol. ahead-of-print No. ahead-of-print.* doi:https://doi.org/10.1108/DLP-10-2020-0106

Kim, J. (2013). Data sharing and its implication for academic libraries. *New Library World 114(11/12)*, 494-506. doi:https://doi.org/10.1108/NLW-06-2013-0051

Kittinger, A. (2020). *Why open that data?* Retrieved from Open that data: https://sites.google.com/view/open-data-oer/home

Machine Readable. (2021). Retrieved from Open Data Handbook: https://opendatahandbook.org/glossary/en/terms/machine-readable/

Molloy, J. (2011). The open knowledge foundataion: Open data means better science. *PLoS Biology, 9(11)*. doi:https://doi.org/10.1371/journal.pbio.1001195

Murray-Rust, P. (2008). Open data in science. *Nature Precedings*. doi:https://doi.org/10.1038/npre.2008.1526.1

National Research Council. (1985). *Sharing research data.* Washington, DC: National Academies Press. doi:https://doi.org/10.17226/2033

NOT-OD-21-014; . (2020, October 29). Retrieved from Supplemental Information to the NIH Policy for Data Management and Sharing: Elements of an NIH Data Management and Sharing Plan: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-014.html

NSF 15-052, Today's data, tomorrow's discoveries. (2015, March 18). Retrieved from National Science Foundation: https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf

OMB. (1999, 09 30). CIRCULAR A-110 REVISED 11/19/93 As Further Amended 9/30/99, Uniform Administrative Requirements for Grants and Agreements With Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations. Office of Management and Budget. Retrieved from https://obamawhitehouse.archives.gov/omb/circulars_a110/

Open Data Commons Licenses. (2021). Retrieved from OpenDataCommons.org: https://opendatacommons.org/licenses/

Pepe, A., Goodman, A., Muench, A., Crosas, M., & Erdmann, C. (2014). How do astronomers share data? Reliability and persistence of datasets linked in AAS publications and a qualitative study of data practices among US astronomers. *PLos ONE 9(8)*, e104798. doi:https://doi.org/10.1371/journal.pone.0104798

Pomerantz, J. (2015). *Metadata.* Cambridge, MA: MIT Press.

Poole, A. (2016). The conceptual landscape of digital curation. *Journal of Documentation, 72(5)*, 961-986.

Porter, J. (2010). A brief history of data sharing in the U.S. Long Term Ecological Research Network. *The Bulletin of the Ecological Society of America, 91*, 14-20. doi:https://doi.org/10.1890/0012-9623-91.1.14

Qureshi, S. (2020). Why data matters for development? Exploring data justice, micro-entrepreneurship, mobile money and financial information. *Information Technology for Development 26(2)*, 201-2013. doi:https://doi-org.libproxy.lib.unc.edu/10.1080/02681102.2020.1736820

Response, O. o. (2015). *Public access to federally funded research: Publications and data.* Retrieved from http://www.phe.gov/Preparedness/planning/science/Pages/AccessPlan.aspx

Rolando, L., Carlson, J., Hswe, P., Parham, S., Westra, B., & Whitmire, A. (2015). Data management plans as a research tool. *Bulletin of the Association for Information Science and Technology, 14(5)*, 43-45.

Sieber, J. (2015, September 9). Data sharing in historical perspective. *Social Science Space*. Retrieved from https://www.socialsciencespace.com/2015/09/data-sharing-in-historical-perspective/

SpringerNature Research Data Policy Types. (n.d.). Retrieved from springernature.com: https://www.springernature.com/de/authors/research-data-policy/data-policy-types/12327096

Strauss, A., & Corbin, J. (1990). *Basics of Qualitative Research.* Sage Publications.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A., Wu, L., Read, E., . . . Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE 6(6)*, e21101. doi:https://doi.org/10.1371/journal.pone.0021101

The Open Definition. (2021). Retrieved from opendefinition.org: https://opendefinition.org/

Thessen, A., & Patterson, D. (2011). Data issues in the life sciences. *Zookeys (150)*, 15-51.

U. S. Department of Energy Public Access Plan. (2014, July 24). Retrieved from energy.gov: https://www.energy.gov/sites/default/files/2014/08/f18/DOE_Public_Access%20Plan_FINAL.pdf

Uhlir, P., & Schroder, P. (2007). Open data from global science. *Data Science Journal, 6*. doi:https://doi.org/10.2481/dsj.6.OD36

UNC Open Access Policy. (2016). Retrieved from unc.policystat.com: https://unc.policystat.com/policy/9376788/latest/

Wikipedia. (2021, March 5). Retrieved from Data Sharing: https://en.wikipedia.org/wiki/Data_sharing

Wilkinson, M. e. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 3*. doi:10.1038/sdata.2016.18

Youngseek, K., & Adler, M. (2015). Social scientists' data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories. *International Journal of Information Management, 35(4)*, 408-418.

Zeng, X., Song, X., Ma, T., Pan, X., Zhou, Y., Hou, Y. Z., . . . Cheng, F. (2020). Repurpose open data to discover theraputics for COVID-19 using deep learning. *Journal of Proteome Research, 19*, 4624-4636. doi:https://doi.org/10.1021/acs.jproteome.0c00316

11 Appendix

Appendix A: Public Funders

agency	plan link	Data plan required	data sharing required	repository provided	limits
Administration for Community Living (under DHS)	https://acl.gov/sites/default/files/about-acl/2017-12/ACLPublicAcccessPlan.pdf	Yes	yes - "The ACL public access plan requires that scientific data generated from ACL/NIDILRR-funded research be publicly available no later than 24 months after an award's end date. The scientific data must be packaged and stored in such a way that enables retrieval and meaningful use by interested parties at no cost. "	preferred repository - "Interuniversity Consortium for Political and Social Research (ICPSR), a unit within the Institute for Social Research at the University of Michigan." Can also choose another public repository	There is an option to state in the data management plan why sharing the data "cannot be justified"
Agency for Healthcare Research and Quality (AHRQ)	https://www.ahrq.gov/funding/policies/publicaccess/index.html	Yes	yes - "AHRQ will require all primary data collected by grant, contract, or intramural research to be submitted to a data repository, unless confidentiality restrictions prevent the data from being made public."	"AHRQ will contract with a commercial repository to accept and manage data submitted by extramural, intramural, and contract researchers. "	"To the extent feasible and consistent with applicable law and policy; Agency mission; resource constraints; U.S. national, homeland, and economic security; and the objectives listed below, digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding should be stored and publicly accessible to search, retrieve, and analyze."
Assistant Secretary for Preparedness and Response (ASPR)	http://www.phe.gov/Preparedness/planning/science/Documents/AccessPlan.pdf	yes	yes - "All ASPR-funded researchers will be required to make the data underlying the conclusions of peer-reviewed scientific research publications freely available in public repositories at the time of initial publication in machine readable formats. "	no	"Digital scientific data that are not in scope for this plan include: • Personally identifiable data • Proprietary trade data • Data related to protecting critical infrastructure • Other data whose release is limited by law, regulation, security requirements, or policy"
Center for Disease Control and Prevention (CDC)	https://www.cdc.gov/grants/additional-requirements/ar-25.html	yes	yes - "Recipients whose terms of award do not include submitting data to CDC are expected to plan and prepare for access to, and archiving/long-term preservation of, collected and/or generated data within the funding period, as set forth below. The final version of a collected and/or generated data set intended for release or sharing should be made available within thirty (30) months after the end of the data collection or generation, except surveillance data that should be made accessible within a year of the end of a collection cycle. In addition, recipients should ensure the quality of data they make accessible and seek to provide the data in a nonproprietary format. If data cannot be made accessible, a justification for not doing so should be provided in the final DMP. "	no	"Data that cannot be de-identified can be provided on request under a data use agreement."
Department of Agriculture	https://www.usda.gov/sites/default/files/documents/USDA-Public-Access-Implementation-Plan.pdf	Yes	yes -"All USDA-funded researchers will be required to comply with USDA's policy for making the digital data underlying the conclusions of peer-reviewed scientific research publications freely available in public repositories in machine readable formats. USDA will ensure that data management plans include clear plans for sharing research data."	yes - Ag data commons	"Alternatively, researchers can explain in their data management plans why long term preservation and access cannot be justified, if applicable. USDA will reserve the right to review and disallow the researcher's argument against long-term preservation and access and require conformance to the access policy for digital scientific data as a condition of funding. "
Department of Defense	https://discover.dtic.mil/wp-content/uploads/2018/06/dod_public_access_plan_feb2015.pdf	Yes	yes - "In accordance with OSTP Memorandum, digitally formatted scientific data resulting from unclassified, publicly releasable research supported wholly or in part by DoD funding should be stored and publicly accessible to search, retrieve, and analyze to the extent feasible and consistent with applicable law and policy; agency mission; resource constraints; and U.S. national, homeland, and economic security."		"publically releaseable, unclassified; For the purposes of this draft proposed plan, data will not be publicly releasable if release would compromise the ability to file for intellectual property protection on any invention arising from the data."
Department of Education	https://ies.ed.gov/funding/datasharing_implementation.asp	yes	yes - "When the Principal Investigator (PI) and the authorized institutional official sign the cover page of the 84.305A, 84.324A, or 84.305N grant application, they will be assuring compliance with IES policy on data sharing as well as other policies and regulations governing research awards. Once the DMP is approved by IES, then the PI and the institution are required to carry it out, and to report progress and problems through the regular reporting channels. Compliance with IES data sharing requirements is expected even though the final dataset may not be completed and prepared for data sharing until after the grant has been completed."	no	"There may be circumstances, such as when a state or district will not allow student data to be released, where investigators will not be able to share their complete data set. However, IES expects primary data collected by the project or extant data obtained from a private source to be shared. In many cases, de-identified data shared through either open or restricted access will be sufficient to meet requirements for protecting the confidentiality of participants."
Department of Energy	https://www.energy.gov/sites/default/files/2014/08/f18/DOE_Public_Access%20Plan_FINAL.pdf	yes	"The merits of the DMPs will be evaluated. This evaluation will take into account the relative values of long-term preservation and access and the associated cost and administrative burden. "	"Open Energy Information Platform (OpenEI), All publicly accessible data on OpenEI will be integrated into the Department of Energy's Enterprise Data Inventory and its Public Data Listing, which can be found on energy.gov/data"	"This section applies to unclassified and otherwise unrestricted digital research data (i.e. digital data required to validate research findings5) - DMPs must protect confidentiality, personal privacy, Personally Identifiable Information, and U.S. national, homeland, and economic security; recognize proprietary interests, business confidential information, and intellectual property rights; avoid significant negative impact on innovation and U.S. competitiveness; and otherwise be consistent with all applicable laws, regulations, and DOE orders and policies."
Department of Homeland Security	https://www.dhs.gov/sites/default/files/publications/DHS%20Public%20Access%20Plan%20-%20FINAL_161229-508.pdf	Yes	yes - "DHS will implement policies that will require, at no more than incremental cost and within a reasonable time, public access without charge to digital data sets gathered in the course of work that meet the Scope criteria above."	yes	"Researchers may explain in their DMPs why long-term preservation and access cannot be justified. - This plan does not require the disclosure of the following categories of information: • Data that contains protected proprietary content (e.g., data collected from Small Business Innovative Research (SBIR)-funded programs). • Data that is subject to International Traffic in Arms Regulations (ITAR) or Export Administration Regulations (EAR). • Classified information. • Sensitive information determined to be For Official Use Only information, Law Enforcement Sensitive Information, Sensitive Security Information (SSI), Sensitive but Unclassified Information, or Controlled Unclassified Information."
Department of Transportation	https://www.bts.gov/ntl/public-access	yes	yes - "This plan to the extent feasible and consistent with applicable law and policy; agency mission; resource constraints; U.S. national, homeland and economic security; and the objectives listed below, require digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding to be stored and publically accessible for search, retrieval, and analysis. This plan requires that awardee(s) and/or the respective Operating Administration ensure Public Access to final research data, subject to the above restrictions and those imposed by data quality and the need to protect national/homeland security, individual privacy, and confidentiality"	no, but will offer assistance in finding a good one	"The intent of the ED PPDG is to increase access to digital data resulting from ED-funded research, including research completed by ED employees as part of their Federal employment, and require that awardees and ED employees ensure public access, at a minimum, to data underlying peerreviewed scholarly publications free of charge to the public, except when otherwise prohibited by law."
Department of Veterans Affairs	https://www.va.gov/ORO/Docs/Guidance/VA_RSCH_DATA_ACCESS_PLAN_07_23_2015.pdf	yes	yes - all data must be shared, but some can be private - "All proposals for VA-funded research must include a data management plan describing the mechanisms for providing public access to the digital data resulting from the research. The plan must specifically include how the final research datasets underlying all publications reporting results of VA-funded research will be made available for discovery, retrieval, and analysis, including which materials will be available in machine readable formats."	no	"All VA-funded researchers will be required to share all digital data underlying the published results from all VA funded research at least under controlled public access mechanisms where privacy, intellectual property, or other concerns preclude open public access. (Effective date: December 31, 2015)"
Food and Drug Administration	https://www.fda.gov/media/90883/download	yes	yes - "To the extent feasible and consistent with applicable law and policy; agency mission; resource constraints; U.S. national, homeland, and economic security; and the objectives listed below, digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding should be stored and publicly accessible to search, retrieve, and analyze."	no - suggest discipline specific	"To the extent feasible and consistent with applicable law and policy; agency mission; resource constraints; U.S. national, homeland, and economic security; and the objectives listed below, digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding should be stored and publicly accessible to search, retrieve, and analyze."
National Institutes of Health - 2013	https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html	yes - if seeking $500,000 or more in direct costs in any year of the project period.	yes - "Starting with the October 1, 2003 receipt date, investigators submitting an NIH application seeking $500,000 or more in direct costs in any single year are expected to include a plan for data sharing or state why data sharing is not possible."		"NIH recognizes that data sharing may be complicated or limited, in some cases, by institutional policies, local IRB rules, as well as local, state and Federal laws and regulations, including the Privacy Rule. As NIH stated in the March 1, 2002 draft data sharing statement (https://grants.nih.gov/grants/guide/notice-files/NOT-OD-02-035.html), the rights and privacy of people who participate in NIH-sponsored research must be protected at all times. Thus, data intended for broader use should be free of identifiers that would permit linkages to individual research participants and variables that could lead to deductive disclosure of the identity of individual subjects. When data sharing is limited, applicants should explain such limitations in their data sharing plans. "
National Institutes of Health - new - effective 2023	https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html	Yes	yes/no - The DMS Policy requires: "Submission of a Data Management and Sharing Plan outlining how scientific data and any accompanying metadata will be managed and shared, taking into account any potential restrictions or limitations.Compliance with the awardee's plan as approved by the NIH ICO. "	no - but they provide a list of potential repositories - https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html	"NIH expects that in drafting Plans, researchers will maximize the appropriate sharing of scientific data, acknowledging certain factors (i.e., legal, ethical, or technical) that may affect the extent to which scientific data are preserved and shared. Any potential limitations on subsequent data use should be communicated to individuals or entities (e.g., data repository managers) that will preserve and share the scientific data. The NIH ICO will assess whether Plans appropriately consider and describe these factors."
National Science Foundation	https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf	Yes	yes - "To the extent feasible and consistent with applicable law and policy; agency mission; resource constraints; U.S. national, homeland, and economic security, digitally formatted scientific data resulting from unclassified research supported wholly or in part by NSF funding should be stored and publicly accessible to search, retrieve, and analyze."	no	"To the extent feasible and consistent with applicable law and policy; agency mission; resource constraints; U.S. national, homeland, and economic security, ; Small Business Innovation Research (SBIR)/Small Business Technology Transfer (STTR) proposals and any other proposal may allow for exceptions for proprietary or otherwise restricted data, including but not limited to personally identifiable information, business confidential information, security, among other concerns outlined in section 4.a. of the OSTP memo."
Smithsonian Institute	https://www.si.edu/content/pdf/about/SmithsonianPublicAccessPlan.pdf	Yes - for digital projects	"This plan requires that an electronic copy or a link to a copy of the final accepted manuscript or the final publication (i.e., version of record) of each covered publication that meets the scope criteria above, as well as its supporting digital research data, shall be submitted to one or more Smithsonian-managed or Smithsonian-approved repositories either twelve months or another negotiated embargo period following official publication date unless a demonstrated special circumstance prevents the covered publication or supporting digital research data from being made publicly available"	Smithsonian-managed or Smithsonian-approved repository	

Appendix B: Private Funders

agency	plan link	final materials public	Data plan required	data sharing required	repository provided	limits
Alfred P. Sloan Foundation	https://sloan.org/storage/app/media/files/application_documents/Sloan-Grant-Proposal-Guidelines-Research-	unclear	Information Products appendix - "The appendix should delineate, in list form, all information products anticipated to be produced by the project."	Information products appendix asks where the data will be stored and other questions about data, doesn't state that data sharing is required though	no	
Bill and Melinda Gates Foundation	https://www.gatesfoundation.org/about/policies-and-resources/open-access-policy	Yes - "final publication should be open access (CC BY 4.0), underlying data and final publication should be accessible and open immediately"	unclear	yes - "All Funded Research including articles accepted for publication shall be available immediately at publication, without any embargo period. Each accepted article must be accompanied by a Data Availability Statement that describes where any primary data, associated metadata, original software, and any additional relevant materials necessary to understand, assess, and replicate the reported study fundings in totality can be found. The Foundation shall require that underlying data supporting the accepted article shall be immediately accessible and open upon article publication. Grantees are encouraged to adhere to the FAIR principles to improve the findability, accessibility, interoperability, and reuse of digital assets."	not provided but suggested (https://www.gatesfoundation.org/about/policies-and-resources/open-access-policy-faq) - "Accepted articles shall be deposited immediately upon publication in PubMed Central (PMC), or in another openly accessible repository, with proper metadata tagging identifying Gates funding. In addition to PMC, grantees are encouraged to deposit their accepted article in a subject specific or institutional repository of their choice."	
Ford Foundation	https://www.fordfoundation.org/the-latest/news/ford-foundation-expands-creative-commons-licensing-for-all-grant-funded-projects/	Yes - licensed as CC BY 4.0, except for confidential material	uncertain	uncertain	n/a	n/a
Hewlett Foundation	https://hewlett.org/about-us/our-policies/	"the Hewlett Foundation now requires that grantees receiving project-based grants—those made for a specific purpose—openly license the final materials created with those grants (reports, videos, white papers, and the like) under the most recent Creative Commons Attribution license. We also will require that the materials be made easily accessible to the public, such as by posting them to a grantee's website. "	unclear	unclear	unclear	"We will not enforce this new requirement thoughtlessly. If our default open license does not make sense for a particular project—such as if the work contains sensitive information or if revenue generated by its sale is critical to an organization's financial well-being—we will work with the grantee to determine the most appropriate license. Our commitment to open licensing is meant to
HHMI – Howard Hughes Medical Inst	https://hhmicdn.blob.core.windows.net/policies/Public-Access-to-Publications	yes - "HHMI strongly encourages all HHMI laboratory heads to publish their original, peer-reviewed research in journals that make publications freely available and downloadable on-line immediately after publication (i.e. open access journals). If a laboratory head chooses to publish an original, peerreviewed research publication on which he or she is a major author in a journal that is not open access, the laboratory head is responsible for ensuring that the publication is freely available and downloadable on-line as soon as reasonably possible after publication, and in any event within twelve months of publication."	n/a	n/a	n/a	n/a
Gordon and Betty Moore Foundation	https://www.moore.org/docs/default-source/Grantee-Resources/data-and-ip-policy-11-2014.pdf?sfvrsn=2	yes - public access to grant outputs	yes - "As part of the grant development process, the foundation may ask prospective grantees to develop a Data Sharing and/or Intellectual Property Plan. In this case, before funding is approved, the foundation and prospective grantee will agree on a plan that reflects the objectives of this policy. Implementation of the plan will be a condition of the grant and incorporated by reference in the grant agreement. The plan should address the topics described in our Data Sharing and Intellectual Property Packet."	sort of - "The foundation's general policy is that Data and Intellectual Property must be managed and disseminated in a manner that leads to the greatest impact. Accordingly, in most cases, Data and Intellectual Property should be owned by the grantee and made available at no cost or, when justified, at a reasonable cost. "		"We recognize there may be circumstances where limited or delayed dissemination of Data, or a more proprietary or revenue-generating approach to Intellectual Property, may be appropriate to protect legitimate interests of the grantee, principal investigators, and research subjects; or because exclusivity may actually lead to greater impact by, for example, providing incentives for future private investment or a sustainability
MacArthur Foundation	https://www.macfound.org/about/our-policies/intellectual-property	yes/maybe -- "Foundation seeks prompt and broad dissemination or availability of the Grant Work Product at minimal cost to the public or, when justified, at a reasonable price."	unclear			
Wellcome	https://wellcome.org/grant-funding/guidance/data-software-materials-management-and-sharing-policy AND https://wellcome.org/grant-funding/guidance/open-access-guidance/open-access-policy	yes - PubMed Central	Outputs management plan - https://wellcome.org/grant-funding/guidance/how-complete-outputs-management-plan	yes - not always public though - "We expect our researchers to maximise the availability of research data, software and materials with as few restrictions as possible. As a minimum, the data underpinning research papers should be made available to other researchers at the time of publication, as well as any original software that is required to view datasets or to replicate analyses. Where research data relates to public health emergencies, researchers must share quality-assured interim and final data as rapidly and widely as possible, and in advance of journal publication. "		
American Heart Association	https://professional.heart.org/en/research-programs/aha-research-policies-and-awardee-hub/open-science-policy-statements-for-aha-funded-research	yes - PubMed Central	yes - "Applicants will be prompted to answer each of the following questions when completing a data plan in the application: What data outputs will the research generate? When will the data be shared? Where will the data be made available? Are any limits to data sharing required?"	yes -" Any factual data that is needed for independent verification of research results must be made freely and publicly available in an AHA-approved repository within 12 months of the end of the funding period (and any no-cost extension)."		"Certain applicants may seek exemption from the Open Data policy. These applicants must submit an opt-out request with the application to explain why the Open Data policy should be waived. Although the applicant may provide other rationale, most opt-out requests fall into one of the following four categories: Human Subject Grounds, Superseding Regulations Grounds,

Appendix C: Publishers

Publisher	link to data policy	require data sharing	in repository	limits	data availability statement	data management plan? or where	about
AAAS Science Journals	https://www.sciencemag.org/authors/science-journals-editorial-policies#research-standards	"All data used in the analysis must be available to any researcher for purposes of reproducing or extending the analysis."	"Data must be available in the paper or deposited in a community special-purpose repository or a general-purpose repository such as Dryad (see Data and Code Deposition)."	"Exceptional circumstances requiring special treatment, such as protection of personal privacy or purchase of datasets from third-party sources, should be discussed with the editor as early as possible (no later than at the manuscript revision stage) and spelled out explicitly in the acknowledgments."	Not specifically stated	Not specifically stated	
American Geophysical Union (AGU)	https://www.agu.org/Publish-with-AGU/Publish/Author-Resources/Policies/Data-policy	"all data necessary to understand, evaluate, replicate, and build upon the reported research must be made available and accessible whenever possible."	"AGU encourages authors to identify and archive their data in approved data centers. If there is no relevant public repository available, and the data are such that they cannot easily be included in a supplement, authors are expected to curate the above data for at least 5 years after publication and provide a transparent process to make the data available to anyone upon request"	"Data sets that are not curated or cannot be reliably made available to anybody requesting data may not be cited in AGU publications. Limitations or restrictions on sharing data must be reported to the Editor for consideration at the time of submission."	"AGU requires an explicit statement in the "Acknowledgments" section of a paper that clarifies how users can access the data from a paper (via supplements, repositories, other sources, etc.) and states any restrictions on access."	"Detailed information describing data or methodology used when the data or methods are new may be presented in one of the following 5 ways: (1) in the main text, (2) in a 'Materials and Methods' section in the manuscript, (3) as Supporting information, (4) as an Appendix, and (5) as a short Companion paper. At the time of first publication online, which is usually a few days after acceptance, any companion paper, and all other references, must be available to other scientists. Papers may be held until companion or referenced papers are	
APA	https://www.apa.org/pubs/journals/resources/data-sharing	"APA's Ethics Code (Section 8.14) instructs researchers to allow other competent professionals access to the data on which their published results are based as long as recipients seek to verify claims through reanalysis—unless confidentiality or legal restrictions prevent sharing."	not required - but they have a repository - https://osf.io/meetings/apa/	Data sharing not required - but encouraged	availability statement indicating whether the data, methods used in the analysis, code, and materials used to conduct the research will be made available to any researcher for purposes of reproducing the results or replicating the procedure. In both the author note and at the end of the method section, either specify where that material will be available or note the ethical or legal reasons for not doing so."		APA encourages data sharing and makes a lot of resources available to learn how to share data but stop short of saying that public data sharing is required, only that data must be made available to other researchers on request
ASTRO	https://www.astro.org/News-and-Publications/Journals/News/Data-Sharing	Data sharing is not required	n/a	n/a	yes - "authors are asked to include a data availability statement with their submitted work. Data availability statements should indicate if the data are being shared and if so, how the data may be accessed."	no	ASTRO does encourage data sharing and provides resources about how to wrte the data availability statement, how to cite a dataset, choose a repository
Elsevier	https://www.elsevier.com/about/policies/research-data	Elsevier encourages and supports data sharing but does not require it.	not required - but they do provide one https://www.elsevier.com/solutions/mendeley-data-platform	n/a	no	no	support but don't require
Emerald Open Research	https://emeraldopenresearch.com/for-authors/data-guidelines	Yes - "Emerald Open Research requires that the source data underlying the results are made available as soon as an article is published."	"Where it is possible to do so, data should be deposited in a stable and recognised open repository under a CC0 license prior to article submission. "	yes - state in the availability statement why it can't be shared; "Data that cannot be shared" includes "Ethical and security considerations", "Data protection issues", "Large data", "Data under license by a third party"	yes - "All articles must include a Data Availability statement, even where there is no data associated with the article. "	advised to link to dataset in the article	
JMIR Publications (Journal of Medical Internet Research)	https://support.jmir.org/hc/en-us/articles/360030832631-What-is-your-data-sharing-policy-	JMIR uses the SpringerNature type 2 data sharing policy and require data sharing with other researchers on request and encourages to share with readers, but doen't require it. JMIR Data uses the SpringerNature type 4 policy which requires data sharing	encouraged - in public repository, main manuscript, or supporting files	n/a	encouraged	n/a	JMIR encourages but doesn't require - uses Springer Nature type 2
PLOS	https://journals.plos.org/plosone/s/data-availability	yes - PLOS journals require authors to make all data necessary to replicate their study's findings publicly available without restriction at the time of publication. When specific legal or ethical restrictions prohibit public sharing of a data set, authors must indicate how others may obtain access to the data."	Sharing in data repository is strongly recommended, can also be shared in supporting information files	Access restrictions allowed for third-party data, human research participant data, other sensitive data	yes	"PLOS encourages authors to prepare DMPs before conducting their research and encourages authors to make those plans available to editors, reviewers and readers who wish to assess them."	
Sage	https://us-sagepub-com.libproxy.lib.unc.edu/en-us/nam/research-data-sharing-policies	Three options: option 1 - encouraged to share data in a repository, data availability statement, cite; option 2 - required to share data in public repository, data avaialbity statement, cite; option 3 - required to share data in public repository, data availablity statement, cite +peer review of research data	depends on option	n/a	depends on option	n/a	
SpringerNature	https://www.springernature.com/de/authors/research-data-policy/data-policy-types/12327096	4 levels: Type 1-Data sharing and data citation is encouraged; Type 2- Data sharing and evidence of data sharing encouraged; Type 3-Data sharing encouraged and statements of data availability required; Type 4-Data sharing, evidence of data sharing and peer review of data required	depends on level	n/a	encouraged or required for 2,3,4	n/a	Springer Nature has made the research data policy texts, unless otherwise stated, available for reuse by the research data community under a Creative Commons attribution license.
Taylor & Francis	https://authorservices-taylorandfrancis-com.libproxy.lib.unc.edu/data-sharing-policies/#tdatapolicies	Options: basic - encourage data sharing and data avaialbility statement; share upon reasonable request; publically availabel; open data - freely available reuse license by any thried party for any lawful purpose - findable, fully accessible; open and fair - "Authors must make their data freely available, under a license allowing re-use by any third party for any lawful purpose. Additionally, data shall meet with FAIR (findable, accessible, interoperable and reusable) standards as established in the relevant subject area."	encouraged - offer advice in finding one	n/a	yes - if journal has a data sharing policy	n/a	
Wiley Journals	https://authorservices.wiley.com/author-resources/Journal-Authors/open-access/data-sharing-citation/data-sharing-policy.html	4 levels: Encourages data sharing - optional - data availability statement, data sharing, perr review; Expects data sharing - require data availability statement, data sharing and peer review optional; mandates data sharing - data availabilty statement required, data saring required, perr review optional; mandataes data sharing and perr reviews data - data availability statement required, data sharing required, perr review required	offers advice	n/a	depends on option	n/a	

Appendix D: Library Science Master's Programs

school	univ. data policy?	univ open access policy	lib/info science school specific data policy	master's paper online repository	masters paper data shared?	data statement requirements	university provided data repository	university research data services
University of Illinois--Urbana-Champaign, Champaign, IL	information security - https://cam.illinois.edu/policies/fo-36/	yes - https://guides.library.illinois.edu/oapolicy	no	yes - https://www.ideals.illinois.edu/handle/2142/8951	no	no https://grad.illinois.edu/thesis/process	yes	yes - https://www.library.illinois.edu/rds/
University of North Carolina--Chapel Hill, Chapel Hill, NC	mostly security - https://research.unc.edu/compliance/data/policies-practice-briefs/	yes - https://unc.policystat.com/policy/9376788/latest/	no	yes - https://cdr.lib.unc.edu/catalog?f%5Baffiliation_label_sim%5D%5B%5D=School+of+Information+and+Library+Science&f%5Bresource_type_sim%5D%5B%5D=Masters+Paper&locale=en	no	no	yes	yes - https://guides.lib.unc.edu/researchdatatoolkit#partners
University of Washington, Seattle, WA	no - privacy	yes - https://www.lib.washington.edu/scholpub/open-access	no	yes - https://grad.uw.edu/for-students-and-post-docs/thesisdissertation/	no - capstone instead but no data sharing	no	yes - https://www.lib.washington.edu/scholpub/scholarly-publishing-services	yes - https://www.lib.washington.edu/dataservices
University of Maryland--College Park, College Park, MD	no - privacy	don't see a specific policy but they do have a respository and resources - https://lib.guides.umd.edu/c.php?g=326747&p=2194003	no	yes _https://drum.lib.umd.edu/	no	no	yes - https://drum.lib.umd.edu/	yes - https://www.lib.umd.edu/data
University of Texas--Austin, Austin, TX	no - security https://security.utexas.edu/policies/irusp	yes for library staff - https://www.lib.utexas.edu/about/policies/open-access-policy-libraries-staff;	no	yes - capstone archive - https://www.ischool.utexas.edu/programs/capstone-archive	no	no - https://gradschool.utexas.edu/academics/theses-and-dissertations	yes - https://www.lib.utexas.edu/research-help-support/research-data-services	yes - https://data.tdl.org/
Rutgers, The State University of New Jersey--New Brunswick, New Brunswick, NJ	ferpa and IT policy	yes - https://www.libraries.rutgers.edu/services-for-researchers/open-access	no - no paper - capstone or e-portfolio	yes - but lis maybe not required to deposit https://rucore.libraries.rutgers.edu/etd/	no	no	https://rucore.libraries.rutgers.edu/research/	https://www.libraries.rutgers.edu/services-for-researchers/data-services
Syracuse University, Syracuse, NY	privacy policy student data,	none found	no	yes - but I don't think lis requires a theis https://surface.syr.edu/thesis/ or proquestETD	no	no	yes but seems to be a fee https://qdr.syr.edu/deposit/process	https://library.syr.edu/services/research.php
University of Michigan--Ann Arbor, Ann Arbor, MI	institutional data resource management	none found	no	yes - deep blue	no	no	yes - https://www.icpsr.umich.edu/	yes - https://www.lib.umich.edu/research-and-scholarship/data-services
Indiana University--Bloomington, Bloomington, IN	administreative data	yes - https://vpfaa.indiana.edu/policies/bl-aca-i24-open-access/index.html	no	yes - https://scholarworks.iu.edu/dspace/	I don't think lis has a paper	no	https://scholarworks.iu.edu/dspace/	https://datamanagement.iu.edu/
University of Tennessee--Knoxville, Knoxville, TN	research data policy - http://research.utk.edu/wp-content/uploads/sites/48/2013/05/policy_research-data1.pdf	yes - https://www.lib.utk.edu/scholar/open-access/open-access-policy-ut-libraries/	no	yes - https://trace.tennessee.edu/utk-grad/	no - https://sis.utk.edu/wp-content/uploads/2020/05/ThesisGuideFinal-rev1.pdf	no	I think they can put data in Trace	https://libguides.utk.edu/dataservices
list of schools from: https://www.usnews.com/best-graduate-schools/top-library-information-science-programs/library-information-science-rankings								