Summer 7-1-2022

Research Data Management Practices And Impacts on Long-term Data Sustainability: An Institutional Exploration

Sarah Elaine Bratt
Syracuse University

Abstract

With the ‘data deluge’ leading to an institutionalized research environment for data management, U.S. academic faculty have increasingly faced pressure to deposit research data into open online data repositories, which, in turn, is engendering a new set of practices to adapt formal mandates to local circumstances. When these practices involve reorganizing workflows to align the goals of local and institutional stakeholders, we might call them ‘data articulations.’ This dissertation uses interviews to establish a grounded understanding of the data articulations behind deposit in 3 studies: (1) a phenomenological study of genomics faculty data management practices; (2) a grounded theory study developing a theory of data deposit as “articulation” work in genomics; and (3) a comparative case study of genomics and social science researchers to identify factors associated with the institutionalization of research data management (RDM).

The findings of this research offer an in-depth understanding of the data management and deposit practices of academic research faculty, and surfaced institutional factors associated with data deposit. Additionally, the studies led to a theoretical framework of data deposit to open research data repositories. The empirical insights into the impacts of institutionalization of RDM and data deposit on long-term data sustainability update our knowledge of the impacts of increasing guidelines for RDM. The work also contributes to the body of data management literature through the development of the “data articulation” framework which can be applied and further validated by future work. In terms of practice, the studies offer recommendations for data policymakers, data repositories, and researchers on defining strategies and initiatives to leverage data reuse and employ computational approaches to support data management and deposit.

Research Data Management Practices and Impacts on Long-Term Data Sustainability: An Institutional Exploration

By

Sarah Bratt

B.A., Ithaca College, 2012
M.S., Syracuse University, 2014

Dissertation

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Information Science and Technology

Syracuse University
July 2022

Dedication

This dissertation is dedicated to Erin Bartolo, a beautiful soul with a fierce intellect.
I draw strength from you every day.

I may not live to see our glory
But I will gladly join the fight
And when our children tell our story
They'll tell the story of tonight

Raise a glass to freedom
Something they can never take away
No matter what they tell you

Raise a glass to the four of us
Tomorrow, there'll be more of us
Telling the story of tonight

They'll tell the story of tonight

Acknowledgements

This dissertation research would not have been possible without the support of many: advisors, friends, colleagues, and family. I would like to express my deep and sincere gratitude to:

- All my interview participants for participating in my dissertation research;
- My advisor, Dr. Jian Qin, whom I admire deeply as a scholar and as a person; for always being patient, available, and kindly guiding and motivating me the doctoral program;
- My committee members, Dr. Kevin Crowston, Dr. Jeff Hemsley, Dr. Bryan Semaan, Dr. Diane Grimes, and internal and external readers, Dr. Carsten Østerlund and Dr. Matthew Bietz, for their insightful input and valuable contributions;
- Faculty mentors and research collaborators, you've taught me so much: Dr. Daniel Acuna, Dr. Rachel Clarke, Dr. Lu Xiao, Dr. Caroline Haythornthwaite, Dr. Jeff Stanton, Dr. Kyle Myers, Dr. Karim Lakhani, Dr. Misha Teplitsky, Dr. Marie Thursby, Dr. Jerry Thursby, and Laboratory for Innovation Science at Harvard (LISH) Fellows;
- All the iSchool and Syracuse University staff, especially Sue Neimer, Shelia Bova-Clifford, Jenifer Pulver, Jennifer Barclay, and Peggy Takach;
- To all my writing group buddies, and those in the program who came to my practice proposal and job talks, including Mahboobeh Harandi, Subhasree Sengupta, Jean-Philip Ramsey, Erin Bartolo, Dr. Huichuan Xia, Leni Krsova, Ania Korsunska, Yiran Duan, Jieun Yeon, Charis Asyante, and so many more;
- Faith Lutheran Church and Lutheran Campus Ministry, for the endorphins of music and laughter;

- For the Syracuse University graduate science policy group (GSPG), especially Vito Iaia, Dr. Ethan Stanifer, Arthur Hernandez and his piano, Dr. Gaye Ceyhan, and Yoda the dog, official GSPG mascot;
- Women in Science & Engineering (WISE) and the Future Professoriate Program (FPP); the Graduate School and Dr. Dan Olsen Bang, for his tireless efforts to assuage my fears and poor grammar throughout my graduate studies;
- La Familia de la Salsa: Roberto Perez, Diane, Samantha, Laura, and everyone at Cardio and Havana Nights; you taught me to embrace my body and the joy of movement;
- Westcott Nation: Recess Coffee, Alto Cinco, Beer Belly, Stout Beard Brewing, Petit Public Library, Dorian's Pizza, Yeti, Gangnam Korean BBQ: third spaces, second homes
- To Dr. Dan Moseson, travel buddy, photography guru, and stellar musician, who has seen me through thick, thin, and so much hot sauce on pizza;
- For my family, for all the video-chats with Grandpa and Elsie, Aunt Wendy, Uncle Tom, Emily, for support, for thoughtfully listening to my research and encouraging me every step of the way;
- To Matt, my rock ('n roll!) and shoulder, my best friend and my chef, my running buddy and explorer of many worlds, I could not have done this without you or the purple futon;
- To my parents, Marilyn and Gary, for their unflagging energy, cartoon-drawing skills, love, all the laughter over thanksgiving dinners and cookie-decorating sessions; for warmly welcoming my friends and colleagues over to dinner.

Table of Contents

CHAPTER 1
PROBLEM STATEMENT

CHAPTER 2
LITERATURE REVIEW

CHAPTER 3

OVERVIEW OF THE RESEARCH DESIGN

CHAPTER 4

STUDY 1

CHAPTER 5

STUDY 2

CHAPTER 6

STUDY 3

CHAPTER 7
DISCUSSION OF FINDINGS & CONCLUSIONS

List of Figures

List of Tables

CHAPTER 1
PROBLEM STATEMENT

1.1 Background

Globally, we have experienced several events in which managing research data has been critical for addressing severe public crises (Álvarez-Machancoses et al., 2020). The SARS-CoV-2 pandemic (COVID-19) vividly illustrated this, demonstrating the crucial role of experimental and clinical genomic data for vaccine development and variant monitoring (Cyranoski, 2021). However, vaccines and variant surveillance were only possible because of the work done to produce and deposit SARS-CoV-2 virus data into open online data repositories, creating a standardized 'big data collection' (Leonelli, 2014a). Standardized data deposit enabled scientists to then search and analyze heterogenous COVID-19 data as a single body of information, even though the data were collected from a diverse range of geographic regions, research sites, and populations (Ankeny & Leonelli, 2015). This data deposit work was done by genomics scientists all over the world who produced, processed, and submitted their data to open online databases such as *GenBank* (Benson et al., 2017) and GISAID (Cyranoski, 2021).

However, the work of data management and deposit into repositories is non-trivial. In fact, during the pandemic, data collection was "patchy" and incomplete (Cyranoski, 2021). Genomics scientists had trouble depositing data because of a lack of resources (Nowakowska et al., 2020), obstacles to coordination (Myers et al., 2020), and other factors that impeded data collection, production, and deposit (Cyranoski, 2021). For example, the SARS-CoV-2 genomes

that the United States, the country with the largest COVID-19 outbreak globally[1], shared on the open online database GISAID was less than 0.3% of the U.S.'s total number of COVID-19 infections (Cyranoski, 2021). The failure to deposit COVID-19 genomic data had severe impacts, including underestimating the virus prevalence, skewing experiments, and affecting the quality of computational models, models which were only as good as the data available to use (Cyranoski, 2021; Leonelli, 2014a). Unless genomics scientists do the work of making data suitable for deposit, it is impossible to develop vaccines, and variant tracking efforts are incomplete or meaningless because of low overall genomic data coverage (Cyranoski, 2021).

The debate on how to support such data management and sharing efforts in U.S. academic research has attracted much attention, given the massive investments in cyberinfrastructure (CI)-enabled science by U.S. federal agencies (*eScience* in the U.K.) (Atkins, 2003; Hey & Trefethen, 2005). In response, data management and sharing in many scientific fields has been increasingly institutionalized – that is, formalized and standardized and become taken-for-granted as an established part of the professional organization of many research fields (Crowston & Qin, 2011). For example, journal publishers require data sharing, federal agencies such as the U.S. National Science Foundation (NSF) mandated a data management and sharing plan, and scientific funding entities – private and public – require data deposit (Crowston & Qin, 2011; Kim, 2013).

A signpost of the CI-enabled scientific milieu is online open research data repositories, institutions for sharing research data. *Data repositories* are constituted of the products, services, and infrastructures for data description, sharing, analysis, and reuse (among other functions) which support data intensive research (Austin et al., 2015). A critical component of this

[1] As of January 2021 (Cyranoski, 2021)

increasing institutionalization of data management and deposit has been efforts to improve the data management practices of researchers, with the goal of enable high-quality data deposit to these online research data repositories. U.S. academic faculty face institutional pressures to deposit their research data to repositories. Among those who deposit data to repositories, faculty are a critical constituency of depositors. Faculty exercise nearly complete control research data management and making data suitable for deposit.

Yet despite the institutional, technical, and human labor support whose aim is to facilitate high-quality data management, supporting faculty data work is not a major focus of investments in CI-enabled science. This threatens to create a gap between data policy and local scientific practice, a gap evidenced by obstacles to data management maturity (Crowston & Qin, 2011) and the lack of conceptual models for explaining how and why researchers deposit their data to open data repositories.

Prior studies in Library & Information Science (LIS), Social Studies of Science (STS) studies, and Computer Supported Cooperative Work (CSCW) have similarly shown faculty data work is non-trivial, and enables data production, deposit, and integration, e.g., data cleaning (e.g., Edwards et al., 2007) and documentation (e.g., Sands, 2017). LIS, STS, and CSCW scholars have also drawn attention to dependencies making data deposit work successful, such as the coordination work scientists do (e.g., Darch et al., 2015), logistical problems preventing data deposit (Akers & Doty, 2013), and the "human infrastructure of cyberinfrastructure" (C. P. Lee et al., 2006).

However, these studies tend not to focus explicitly on the institutional factors that constrain and/or facilitate data management and deposit practices. If the grand visions of cyberinfrastructure-enabled data sharing are to advance, researchers' data management and data

deposit practices need to be supported by the very institutions who aim to facilitate and accelerate data-intensive science. If not, research data is lost, processes remain inefficient, undermining trust in science, and taxpayer dollars aimed to address critical contemporary issues from climate change and COVID-19 to cultural heritage preservation and the increasingly *datafied* contexts of our everyday lives.

1.2 Motivation

The literature has started to investigate data sharing as a practice, "engaging the ongoing activities of researchers" (e.g., Widmalm, 2016). These studies have made valuable contributions in the forms of ethnographic description and survey analysis. They led to a greater attention towards the diverse range of technologies, actors, and values "behind and beyond the datum itself" (Neang et al., 2020). However, where the literature does engage the data practices behind data management and deposit explicitly, it uses practice as an explanation or unit of analysis, and is largely limited to identifying the factors that contribute to when or whether data sharing and reuse will occur or how they can be incentivized (e.g., Tenopir et al., 2011). Of the studies, many outline enabling factors and obstacles to data deposit, including "...documentation, access, collaboration politics, standardization efforts, disciplinary differences, and concerns around scooping … that complicate data exchange" (Neang, 2021). Less attention has been directed toward the data deposit work that constitutes the intersection of data practices and institutional factors which make data deposit successful in U.S. academic faculty research.

To address this gap, I adopt the theoretical frameworks of "institutionalization" and "articulation work." *Institutionalization* helps us analyze how data mandates and other instruments of institutionalization interact with faculty research data management (RDM) and deposit practices. The data work of faculty is embedded in their institutional contexts. That is,

the way faculty's professional and academic institution is set up shapes what options are available for making data management and deposit decisions (Kim, 2013). Put another way, faculty make decisions as members of professional associations, academic disciplines, and academic universities and as the recipients of grant funding and as authors who submit to journals (Kim & Stanton, 2016). At the same time, institutional directives like data mandates are not just a policy imposed upon faculty – policy implies an infrastructure, one intended to assure the implementation of the policies, mandates, and norms.

Institutional theory can provide insight about how social actors are influenced by institutional pressures from the institutional environment. While traditional unit of analysis in institutional theory was the organizational level, neo-institutional theory extends its scope to a variety of social actors, including organizations and individuals (W. R. Scott, 2013). Neo-institutional theory posits that institutional environments including institutional rules, norms, and culture influence individuals' perceptions, behaviors, and attitudes (George et al., 2006; Lounsbury et al., 2021; Tolbert, 1985; Tolbert & Zucker, 1983) . Contemporary perspectives on institutional theory also consider individual beliefs concerning proper social behavior and, specifically, when those beliefs arise from organizational rules, structures, and practices (Barley & Tolbert, 1997; Daniels et al., 2002; Duxbury & Haines Jr, 1991; Kim & Adler, 2015). Thus, the institutionalization framework enables us to understand the ways that data practices are formalized and standardized and interact with faculty's data work and the norms and beliefs about appropriate behavior for RDM and deposit.

Yet, faulty must do "articulation work" to meet institutional pressures. Because institutional mandates and data policies cannot always specify what specific actions will be needed in all local circumstances, faculty must align and tailor the mandates to "a set of

implementation conditions that cannot be fully specified ahead of time" (Gerson & Star, 1986). Since no centralized authority or institutional policymaking body can possibly anticipate all the contingencies that might arise locally, faculty always have some discretion in deciding how the policies are implemented. No matter how detailed the requirements are, they must be customized by making "local adjustments that made the work possible in practice" (Gerson & Star, 1986). These adjustments to implement the institutional policies, norms, and pressures are called "articulation work.".

Articulation work consists of "all the tasks needed to coordinate a particular task, including scheduling subtasks, recovering from errors, and assembling resources" to align levels of work organization (Gerson & Star, 1986). Articulation makes it possible to have standardized processes for research data management, for deposit work and its products, as captured in mandates, instructional manuals, data training curriculum, documentation, and databases. The work of articulating RDM and data deposit involves making sense of data at every stage of the research cycle, an intellectual process of "crunching the data," enacted within organizational processes of aligning levels of work organization, to lay down the way for analysis.

The "articulation" framing allows us to see what work enables data deposit because it constructs data management and deposit as processes of aligning the concerns of field-level institutional goals and lab-level research practices to make research data deposit 'do-able' (Fujimura, 1987). When articulation work is "deleted in idealized representations of that work…the resulting task descriptions can only be uneasily superimposed on the flow of work" (Gerson & Star, 1986). The current proliferation of 'data lifecycle models' to describe data workflows reflects this state of affairs (A. M. Cox & Tam, 2018; S. T. Kowalczyk, 2018).

Without an understanding of articulation, the gap between requirements for data policy and the actual data work of faculty will remain inaccessible to scholarly analysis.

Together, the “institutionalization” and “articulation” perspectives help to develop evidence-based policy and responsive system design for research data management (RDM). To understand the institutionalization of RDM and the articulation it may engender both practically and theoretically, we need to have detailed empirical analysis of faculty data management and deposit work in context. Here, I argue these perspectives offer three contributions:

1. As more institutions start RDM projects and initiatives, organizations will increasingly demand guidelines for research data management. While some fields boast mature RDM infrastructure, others are still developing. Those which lack guidance on how and whether to institutionalize data management practices can potentially learn from disciplines with mature institutional infrastructures for RDM.
2. Institutionalizing a practice can increase workflow efficiency and promote transparency but requires articulation work. In the context of academic RDM, institutionalizing data work can support long-term data sustainability. Thus, we can develop evidence-based policy to support long-term research data sustainability by adopting both institutionalization and articulation perspectives.
3. Institutionalization also makes practices legitimate and mandated. In taking the institutionalization perspective for RDM, we can assure the necessary resources and infrastructure exist for data deposit to occur.

In this dissertation, I conduct a qualitative empirical study of the faculty data work situation, emphasizing the articulation- institution dynamics. The dissertation overall addresses

the need to better understand how articulation practices and institutional factors intersect to enable and constrain research data management (RDM) and deposit.

1.3 Research Goal & Research Questions

This dissertation seeks to add to the small but growing literature examining how scientists manage and deposit research data in open online research repositories. The goal aligns with the emerging group of LIS, CSCW, and STS researchers who aim to "move from the delineation of factors or barriers towards the particular kinds of work that researchers do in developing practices for overcoming barriers and bringing their concerns into alignment" (Neang et al., 2020). These studies explore the gap by asking: What work do scientists do to make data suitable for deposit? I suspect this work impacts on long-term research data sustainability, including data quality and *whether* data are deposited at all.

Stated succinctly: This dissertation seeks to develop a conceptual framework for researchers' efforts to deposit data in an increasingly institutionalized research environment. I argue faculty face pressures to produce and deposit data because the current research environment values and incentivizes practices associated with long-term research data sustainability. Paying attention to these data management and deposit practices matters for computational model quality, research data systems design, workflow efficiency, and faculty development outcomes, and ultimately long-term research data sustainability.

The central question of this dissertation is: *How do scientists manage and deposit data in increasingly institutionalized research environments?* To address this central question, there are four guiding research questions across 3 studies (*Dissertation Outline*, **Table 1**).

Table 1: Research Questions and Dissertation Outline

Dissertation Outline	
Method & Central Question(s)	**Purpose & Participant Population**
Study 1	
Phenomenological study *RQ1: What are the experiences of data deposit for genomics faculty?*	*Purpose:* To surface the experiences and meaning of *data deposit* for faculty within data-intensive genomics research environments **Participants**: Molecular biology faculty (n = 12), U.S. R1 academic institutions
Study 2	
Grounded theory-inspired study *RQ2: What faculty data practices make data deposit 'do-able'?*	*Purpose:* To explain the process *data deposit*, building on the experiences of faculty's work making data suitable for deposit **Participants**: Genomics faculty (n = 18), U.S. R1 academic institutions
Study 3	
Comparative case study *RQ3: What institutional factors are associated with "articulation" of data management and deposit?* *RQ4: What are impacts of "articulation" on long-term research data sustainability?*	*Purpose:* Identify the institutional factors associated with the articulation work involved in deposit data. Identify the impacts of articulation and institutionalization on long-term sustainability. Test and elaborate the "articulation" framework in broader population, by comparing high- and low-institutionalization data deposit contexts. **Participants:** Genomics faculty (n = 21) and social science faculty who deposited research data to ICPSR (n = 15), U.S. academic institutions

The overarching design of this research is to take a sequential qualitative approach in three studies: Study 1 is a phenomenological study. Study 2 is a grounded theory study. Study 3 is a comparative case study. The purpose of the phenomenological study (Study 1) was to explore the experiences of genomics scientists at R1 U.S. research universities in depositing data. The purpose of the grounded theory-inspired study (Study 2) was to develop a theory to explain the activities that enable data deposit, based on the phenomenological study findings. The purpose of the comparative case study (Study 3) was to verify and elaborate the theoretical framework developed in the grounded theory study in a broader group.

1.4 Expected Contributions

This dissertation makes theoretical, methodological, and practice-based contributions. The primary contribution of this dissertation is to address the emerging issue of how to support stakeholders in the institutionalization of data deposit. is an identification of the factors To do this, the study applies and develops the theory of *articulation work* in the context of institutionalized data deposit in genomics and the social sciences.

Theoretical contributions: This study developed the "data articulation framework" to explain how scientists deposit data to open research data repositories. With this understanding, researchers can better isolate variables and develop models to analyze requirements for supporting data deposit.

Methodological contribution: Using metadata from GenBank and ICSPR as a critical-incident technique during the qualitative interviews to offers a methodological contribution to ameliorating self-report inaccuracies. Second, the request of documents used for RDM in the scientists' lab augments approaches document analysis to trace the articulation work of scientists, by leveraging on trace and virtual ethnography-inspired approaches.

Practice-based contributions: By examining institutional contexts of data deposit in genomics and social science using a qualitative approach, we can better understand the conceptions and work practices of scientists that make data deposit possible in U.S. academic research. Science policymakers can plan interventions and allocate resources to prevent data loss to for long-term research data sustainability and professional organizations can assist with faculty development. More broadly, examining the work of data deposit also disabuses us from 'big data' imaginaries by understanding the faculty work involved in data management to meet emerging institutional demands.

1.5 Key Terms

This study is built on four key concepts: *digital scholarship*, *research data management* (RDM), *data deposit*, *institutionalization*, and *articulation*. While these terms have already been introduced above, they are provided again here for both clarity and convenience. The definitions below are not meant to include every sense of the concepts; instead, they represent the meaning of the terms most relevant for the study. These definitions serve to specify the technical sense(s) of the key terms used throughout this document and in future work.

The key terms and definitions are drawn primarily from the scholarly communication literature but are culled from relevant conceptualization in knowledge management, economics, and sociology, among others (reviewed in Ch. II). The terms are described in order of specificity, starting with the general area of *digital scholarship* and *research data management* as a category within, then narrow to focus on the phenomenon of *data deposit* through the lens of the *institutionalization* of RDM and data deposit as a type of *articulation* work.

1.5.1 Digital Scholarship

Digital scholarship is defined as "the use of digital evidence, methods of inquiry, research, publication and preservation to achieve scholarly and research goals" (Ayers, 2004; Garnett & Ecclesfield, 2011; Raffaghelli et al., 2016; Stewart, 2015; Trinkle & Andersen, 2015). The definition used here draws primarily from literature in faculty development and education, a discourse that tends to use Boyer's model of scholarship (1997) adapted to digital contexts and environments. *Digital scholarship* is used in the context of this document for two chief purposes. First, the concept is used to signal the study is situated in the dialogue in digital scholarship, a dialogue occurring primarily in education and faculty development, library and information science (LIS) studies of cyberinfrastructure-enabled science, and digital humanities.

Second, *digital scholarship* is used to foreground digital technologies as an overlooked but important mediator of RDM best practices. Because digital scholarship focuses on the characteristics and affordances digital tools intersected with academic science roles and responsibilities, it enables an analysis the impact of digital environments and tools on the institutionalization of faculty's RDM practices. *Digital scholarship* is a useful perspective for bringing out the RDM processes and practices impacted by digital technologies mediating the institutionalization of RDM.

1.5.2 Data Deposit

Data deposit is defined as the actions, practices, and processes of submitting data to an institution, database, or repository. By this definition, 'data submission' is synonymous with 'data deposit' (Borgman, 2015). Data deposit includes the work of producing and making data suitable to be submitted to an online data repository. By this definition, *data deposit* is not a single terminal event in which a scientist goes to a repository and submits their data. Rather, the concept of data deposit encompasses the data production, sharing, deposit, and other work that *enables* and *constitutes* dataset submission to a repository. Deposit is often required by institutions such as funders or journal publishers. In this study, I focus on data deposit exclusively to open online research data repositories. For example, federal databases for genomic structure and function include GenBank, EMBL, GEO, and model organism repositories. Institutional data repositories also include *Dataverse* and the Inter-University Consortium for Political and Social Research (ICPSR).

1.5.3 Institutionalization

Institutionalization spans a variety of still-evolving analytic and theoretical traditions. I draw from neo-institutionalization to define the extent to which data management and deposit has been institutionalized in U.S. faculty research groups (DiMaggio & Powell, 2000).

Specifically, in this document I draw from the three pillars of institutionalization: Regulative, Normative, and Cultural-Cognitive (Greenwood et al., 2017). I operationalize the extent to which data management and deposit is institutionalized with the capability maturity model (CMM) for research data management (RDM) developed by (Crowston & Qin, 2011).

There are multiple levels which the unit of analysis of a study can capture. In interviewing faculty about their data practices (unit of observation) to understand the extent to which they are institutionalized among their research group (unit of analysis), I focus on the "micro-foundations" of the institutionalization of data management and deposit. The on micro-foundations of data management are formalized as the attitudes, behaviors, and agency exercised by individuals embedded in the macro-institutional context of U.S. CI-enabled science. I argue with Ribes (2019) that this period of CI in U.S. science and technology policy began in the early 21st century by seeded by the NSF *Atkin's Report* (Atkins, 2003).

1.5.4 Articulation

Articulation is defined by adapting Joan Fujimura's (1987) alignment concepts to uncover the articulation work of data deposit. Fujimura conceptualized the do-ability of scientific problems as the alignment of three levels of work organization (experimental, laboratory, and social world) and two types of work (production and articulation) (Fujimura, 1987). In this conceptualization, there are three levels of work organization scientists need to bring into alignment to make scientific problems feasible. The three project levels of work organization are: (1) experiment level in which a set of tasks are performed in the laboratory, the (2) laboratory level as a collection of several experiments and other tasks like purchasing laboratory equipment (e.g., an ultracentrifuge) and the (3) social world level, that is, the broader social milieu in which experiments and laboratories are situated (Gerson, 1983; Strauss, 1978).

By this definition, scientists accomplish alignment by articulating tasks across project levels. Articulating means ‘considering, collecting, coordinating, and integrating’ between the levels of work organization Fujimura, 1987: p. 258). In other words, scientists make problems doable through the practices of "organizing and reorganizing work” (Fujimura, 1987).

1.6 Overview of Research Design

This dissertation is designed as a sequential study using a qualitative approach in three studies. Each study has a purpose which contributes to the overall goal of addressing the empirical and theoretical gap of the institutionalization of data and faculty data practices in the context of managing and depositing research data to an open research data repository.

The purpose of the phenomenological study (study 1) was exploratory. The goal was to surface the practices, experiences, and meaning of data deposit for faculty in data-intensive genomics research environments. The central question of the phenomenological study was: What work does it take and what does it mean in genomics research to deposit data? This study found RDM has become more institutionalized, and that faculty often must reorganize their workflows to accommodate directives, mandates, and cultural pressures to deposit data.

The purpose of the grounded theory-inspired study (study 2) was to develop theory to explains these experiences of depositing data by genomics researchers, because theories of the social processes that enable data deposit do not exist for the genomics population. To address this gap, the central question of the grounded theory study is: What is a theory that explains the process of data deposit? This study found faculty engage in articulation work in response to institutional pressures to deposit data. The processes include setting checkpoints or ‘thresholds’ for data deposit to ensure data is suitable and contingencies met. We also found some outcomes

of articulation are aligned with the goals of long-term research data sustainability including ensuring data quality, integrity, and completeness.

The purpose of the comparative case study (study 3) was to determine the institutional factors associated with articulation, applying the framework to a broader population (i.e., genomics and social science data deposit). The central questions are: What factors are associated with articulation work in 'big science' (highly institutionalized data deposit) and 'little science' (low data deposit institutionalization) research data management (RDM) and deposit contexts? (2) How does the institutionalization of research data management (RDM) and deposit impact articulation? and (3) What impacts does the institutionalization of data deposit have on long-term research data sustainability? The findings of the study reveal factors associated with articulation include the absence institutional buy-in at the university level, and a lack of clear guidelines or precedence for data management in the social science field. We also found that faculty in the social science face an institutional gap between supportive infrastructure for data deposit by appropriating or developing resources to assist with data deposit (e.g., lab handbooks, data analysis templates).

These three studies all contribute to the dissertation's overarching goal to a) identify the factors associated with 'articulating data institutionalization' in big science and little science fields; and b) identify the impacts of articulation on long-term research data sustainability. Genomics represents a 'big science' field with mature data institutionalization (e.g., data deposit to GenBank). Sociology and political sciences represent a 'little science' field with less institutionalized data deposit practices (e.g., data deposit to ICPSR). Note that Inter-university Consortium for Political and Social Research (ICPSR) *is* an institution, seeming to indicate a high level of institutionalization. Although ICPSR is literally an institution, the mere presence of

an institution for data deposit does not provide evidence for institutionalization of data deposit. Such an observation conflates a colloquial definition of institutionalization – which considers only the regulative pillar of institutions (Scott, 2013) – as a necessary and sufficient condition for "institutionalization." However, with the well-established definitions in the literature, institutionalization of data deposit here encompasses the regulative, normative, and cultural-cognitive pillars of institutions to establish data deposit as taken-for-granted and legitimated (Scott, 2013).

In this study, then, I operationalize *institutionalized data deposit* as the extent to which faculty's attitudes about and practices related to data deposit show evidence of the "indicators of institutionalization" according to Scott's (2013) three pillars of institutions. I draw from this definition to develop a select subset of the indicators as operational measures for *institutionalized data deposit*. For example, indicators of pressure from the *normative pillar* include the extent to which social acceptability of data deposit exerts pressure on faculty deposit; (2) the *regulative pillar* is indicated by the extent to which data policies exist and pressures faculty to deposit data and (3) the *cultural-cognitive* is indicated by the taken-for-granted nature of data deposit practices of faculty.

The social sciences data deposit practices are less institutionalized in terms of the extent to which data deposit is taken-for-granted and legitimated across Scott's three pillars of institutions (normative, cultural cognitive, and regulative). Although ICSPR is an institution, many of the key indicators of the institutionalization of data deposit are lacking according to the key indicators of institutionalization. In summary, the purpose of selecting social sciences as a field to study, was because data management is fairly localized and field level of

institutionalization is lacking, as indicated by prior literature (R. G. Curty, 2015; Jeng et al., 2017; Jiang et al., 2021; Mozersky et al., 2021).

The study design focuses on data deposit to data repositories because they are a signpost of data institutionalization and rich site to study how faculty adapt data policy to local circumstances through articulation activities.

The broad design of the study is qualitative and leverages multiple qualitative research approaches, including naturalistic inquiry and grounded theory. Qualitative investigations are valuable because they can provide detailed views of the participants (e.g., faculty) in their own words, complex analyses of multiple perspectives, and specific contexts (e.g., of different academic labs), and the institutional pressures that shape genomics researchers' experiences with data deposit. Moreover, qualitative inquiry offers the opportunity to involve scientists as co-researchers, a feature that can enhance the salience of participant views uncontaminated by institutional, administrative, or policy perspectives (Creswell & Poth, 2016; Stephan, 2012b). The overarching research design are further detailed in Chapter 3.

1.7 Chapter Summary

This dissertation study of faculty data practices is organized into 7 chapters. Chapter 1 introduced the background of the study and presented motivating literature and the study rationale, research questions and expected contributions, as well as the definitions of key terms. Chapter 2 provides a targeted review of the literature relevant to all three studies of the study. Chapter 3 describes the research design of the dissertation and common components across the studies (e.g., data description, data storage protocol). Chapter 4 reports Study 1, an exploratory study of faculty data practices in genomics. Chapter 5 reports Study 2, which develops a theoretical model of articulation work to explain the mechanisms of data deposit and

institutionalization found in the first study (in prep for submission to STHV). Chapter 6 presents Study 3, which elaborates the theoretical model developed in Study 2 by providing a report of a comparative study of social and political science faculty who deposit to ICPSR analyzed using the data articulation model (submitted to ASIST 2022). Chapter 7 is the discussion, limitations, positionality statement, and conclusion of the studies, tying the three studies together to discuss the impacts of institutionalization of data deposit for on faculty workflows and the implications for long-term research data sustainability.

The studies comprising this dissertation are timely because many disciplines are moving toward making their data public (e.g., because of open access journals), increasingly developing research data management (RDM) guidelines. As we move toward that model, we are going to see a lot of researchers trying to understand this process which they have not before. They will have to learn as they and learn how to deal with this formalization, and the domain-specific implementations of RDM. By examining multiple genomics scientific work contexts using qualitative approaches and involving scientists as co-researchers, we can better understand the conceptions and work practices of scientists that make data deposit possible in U.S. academic research. With this understanding, researchers can better isolate variables and develop models about data deposit. Science policymakers can plan interventions and allocate resources to prevent data loss and enable effective data deposit workflows for long-term research data sustainability. Academic institutions and organizations can support faculty in managing and depositing their research to promote FAIR data by facilitating data curation and sharing.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

Studies of research data management (RDM) draw from a variety of fields and have served to increase our understanding and awareness of research data practices in the digital age, highlighting a changing scholarly research landscape. They direct our attention to research data as critical assets that must be managed to build the scholarly research infrastructure. These studies indicate organizations are increasingly demanding guidelines for data management, which has shaped researchers' data management practices, introducing an institutional presence.

Despite this understanding, little work within research on data management practices in research settings has been framed explicitly as a process of institutionalization, nor have studies examined the implications of this institutionalization for long-term research data sustainability. Institutionalization is the process whereby practices and/or values become taken-for-granted and legitimate. Institutionalization is enacted through activities that establish, maintain, adapt, and transform rules and behaviors and impacts how individuals and entities interact.

As such, research data management and deposit work practices in U.S. academic research harbor the potential to serve as a rich site for studying how institutionalization impacts data practices and the impacts on long-term data sustainability. However, this potential is yet to be fully explored, even in domains with mature institutional guidelines for data management and

deposit (e.g., '*big science'* disciplines like genomics) or in fields with rising needs for data management guidelines (e.g., '*little science'*[2] disciplines, e.g., ecology).

The genomics domain has both a strong history of institutional norms and a set of influential and widely shared standards and technological artifacts encoding them. An institutional analysis of their practices harbors opportunities to reveal best practices and key challenges to inform policy interventions, develop theory to explain how researchers and institutional instruments interact, and generate information system design requirements. This is the purpose of the present study, and this literature review is intended to provide the context for such work.

To do this, the chapter places the present study with a broad foundation in scholarly communication, and the intersection of two bodies of research: the study of research data management (RDM) and the study of the institutionalization of RDM. Each of the corresponding sections provides an overview of relevant streams of research, including specific topical areas, influential theories, methodological approaches, and major findings. Following these, brief summaries serve to highlight the connections between these areas of research and posits the intersection as a starting point for the current study. The literature covered here is broadly applicable to all three papers in this dissertation. However, each study includes additional literature specific to its focus and emergent findings. For instance, a major finding of the study is that *articulation work* is needed to enact the institutionalization of data management. In this chapter, articulation work is only briefly covered in the sections where scholarly communication

[2] When I refer to 'big science' and 'little science,' I draw from the definition put forth by (Price, 1963) of well-resourced, centralized science that do 'moonshot' projects adapted by scholars to refer to data (e.g., Borgman, 2015). *Big science* refers to large teams of well-resourced science with centralized organizational structures and standardized methods and often 'big data' (e.g., the Laser Infrared Gravitational Observatory (LIGO) project). *Little science* refers to smaller groups, shorter projects timespans, and 'little' heterogeneous data (Crowston & Qin, 2011).

studies have employed it to understand research data management practices and their institutionalization. Theories of articulation are covered thoroughly in Study 2, given that it emerged in that study.

Throughout the chapter, interdisciplinary literature is drawn from across relevant studies spanning scholarly communication studies of faculty digital scholarship practices, science & technology studies (STS), computer supported cooperative work (CSCW), law and public policy, economics, organizational theory, and education and faculty development literature. Each section of the chapter includes core definitions, empirical findings, influential theoretical work, and relevant methods. The chapter concludes with a synthesis of the literature and discussion that highlights opportunities for future research, which the first study (paper 1) addresses.

2.2 Scholarly Communication & Digital Scholarship

This section describes the digital scholarship environment and some of its most important impacts that are shaping the research data management environment. Understanding digital scholarship will assist in setting the context for the following studies which are aimed at examining the data management efforts of U.S. faculty in institutionalized environments.

Scholarship has been at the nexus of the fourth paradigm, an era characterized by more computation, collaboration, and data-intensive activity than previous eras (i.e., experimental, theoretical, and computational eras) (J. Gray, 2009; Hey et al., 2009; Szalay & Blakeley, 2009). *Scholarly communication* has been traditionally defined as the formal and informal channels of exchanging meaning between communities of research practices (Borgman, 1990; Meadows, 1997). As a phenomenon, scholarly communication manifests as conversations, documents, processes, and artifacts, as well as their interactions.

As scholarly communication became increasingly conducted in digital environments, the area of *digital scholarship* emerged to examine how digital tools and environments impact scholarly processes, practices, and products (Ayers, 2004; Raffaghelli, 2017), such as the use of born-digital evidence, methodological approaches, publishing, and archiving in research (Rumsey, 2017). *Cyberinfrastructure (CI)-enabled science* is related term for digital scholarship, and scholars in the area have likewise studied how it extends traditional understandings of the relationship between information, documents, and social structures in science (Cronin & Sugimoto, 2014; Ni et al., 2013; Priem, 2014; West et al., 2013). *Cyberinfrastructure* is defined as the integrated system of networks, hardware, software, and "middleware" and is designed to enable a variety of data acquisition, management, storage mining, and other activities over the Internet (Atkins, 2003; Gold, 2007). The Atkin's report argued cyberinfrastructure is an essential component of an information economy: "If infrastructure is required for an industrial economy, then we could say that cyberinfrastructure is required for a knowledge economy" (Atkins, 2003).

Digital scholarship and CI-enabled science has been particularly consequential for research data practices, and data management and deposit specifically, shifting both how research is conducted and what products and processes are of value. By making it easier to share, copy, and cite data, digital technologies for have amplified the value of data. For example, sharing and copying data enables the creation of 'big data collections,' bodies of information amenable to large-scale computational approaches (e.g., machine learning and artificial intelligence (AI), which have proven to yield novel scientific insights, e.g., by amassing larger population samples or accumulating more comprehensive longitudinal data. Digital object identifiers (DOIs) and hyperlinking are other exemplary digital affordances which significantly broadened the landscape of scholarship by amplifying the value of data by creating permanence.

Digital scholarship ushered in new incentives and constraints that define the *de facto* decision space in which researchers make choices and conduct data work. As a result, digital scholarship has initiated the introduction of new types of scientific products. While publications remain the preeminent research output, there has been a rise in the institutional recognition of the value of data sharing and management. As a result, the scholarly landscape has been articulated by new incentivize structures promoting data management, sharing, and reuse.

Yet despite the changes digital scholarship can seem to invite, researchers' practice still often reflects the priorities of traditional institutional norms. Data are still not part of the crediting culture. For example, tenure and promotion (T&P) evaluations are not often inclusive of data. Scholarly practice today exists in a field of tensions between academic orthodoxy or digital openness (Esposito, 2013) because digital scholarship complicates traditional models of scholarship due to not only technological changes but also cultural shifts, most notably, the open science movement, the breakdown of centralized expertise (Collins & Evans, 2008), and "networked participatory" scholarship (Stewart, 2015). Granted, there have been some significant institutional changes. For example, the open science movement provided an "unprecedented opportunity" for libraries to reflect on their practices and services (J. Cox, 2016; Darch et al., 2020b). Centers for Digital Scholarship now proliferate in U.S. university libraries, offering consulting services for data management (Cox, 2016).

In summary, traditional scholarly communication systems still dominate much of the incentives and channels of communication. However, the landscape is shifting with digital technologies, distributed work, and datasets as end rather than mere intermediary products of research. Within this shifting research environment, a key area implicated has been research data management (RDM) and data practices in academic research.

2.3 Research Data Management (RDM)

The use and management of data is one a core component of scientific work. This is referred to as *data scholarship,* that is, the use of representations used as forms of evidence within the scientific knowledge production process (Borgman 2015). Data scholarship activities are commonly performed with digital environments and tools, making it a category of digital scholarship (Raffaghelli et al., 2016). Research data management (RDM) is a data scholarship activity and shared concern of researchers across disciplines from STEM fields to the social sciences and humanities. Though they share RDM concerns, the nature of the data in each field and project varies widely, impacting the form RDM takes. The disciplinary or lab culture also impacts the performance of RDM and the level of process visibility and management.

In this section, the theoretical foundations of research data management are reviewed. First, a broad definition of *scientific research data* is provided then narrows to the context of data intensive biosciences research with a focus on genetics. The chapter then presents literature in RDM lifecycles and the workforces and expertise required for managing research data management, with a focus on data intensive RDM and the roles and responsibilities of faculty specific to the context of small group data intensive genetics projects in U.S. research institutions. The section concludes with a section summary and synthesis of the literature.

2.3.1 Research Data

Western science is premised on evidence-based processes. Whether the goal of the research is theoretical or applied, the methods of inquiry are grounded in justifying conclusions using evidence. Thus, a central feature of modern scientific inquiry is the collection and analysis of data. While the use of data is widespread and there is a general colloquial definition, a technical definition of research data remains ambiguous and a non-universal concept (Borgman, 2015;

Borgman et al., 2015; Renear et al., 2010). However, the recent spotlight on data and data intensive science have led to finer grained and more precise conceptualizations of research data.

Data are defined as the representations or inscriptions constituting evidence to support an argument, propositional knowledge claim, or premise (Borgman, 2015). This conceptualization of data draws from theories of data that emphasize their materiality, the importance of their context, and their socially constructed meaning. Rather than '*what* are data,' this interpretivist perspective (Pickard, 2013) of *data* proposes an alternative construction the question; rather, "*when* are data" (Engeström 1990). There are ontological assumptions implied by the question 'what are data?' which interpretivist and post-positivist scholars assumption this definition point out that it is "often better to have the right data than more data" ' (Borgman, 2015).

Even among collaborative research teams, what constitute 'data' are construed differently: "What are data to the science teams may be context to the technology teams…" (Borgman et al., 2012). There are longstanding debates in information science and socio-materiality, among others, on the definitions of *data* and its relationship to related communication concepts, such as artifacts, documents, information, knowledge, and even 'wisdom' (Buckland, 1991; Furner, 2004; Jennex, 2009; Tuomi, 1999). For example, the definition provided by the *Open Archival Information Systems* (OAIS) defines *data* as distinct from *information*: "[data] are reinterpretable representation[s] of information in a formalized manner suitable for communication, interpretation, or processing" (Lee, 2010).

Scientific research data is often illustrated with lists of examples, rather than comprehensive definitions. The OAIS provides examples of data including "a sequence of bits, a table of numbers, characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen" (C. A. Lee, 2010). Social science data repositories (e.g. the Inter-

University Consortium for Political and Social Research (ICPSR) defines data similarly, listing examples that include artifacts and documents: “transcripts, audiovisual material, Web sites, geospatial data, biomedical data, and digital video” (Vardigan & Whiteman, 2007).

To address the some of the ambiguities for practical purposes of data management, information practitioners, scholarly communication theorists, and science studies scholars have emphasized the importance of context in managing data. One way of capturing context is using *metadata*. *Metadata* enables a more exact description and fine-grained typologies of scientific research data, even as data contexts dynamically change and are interpreted differentially between disciplines and among project team members. *Metadata* is “documentation, descriptions, and annotations created and used to manage, discover, access, use, share, and preserve informational resources” (Mayernik et al., 2011). According to Zeng and Qin (2008), metadata is “structured data” that are “encoded in various formats, governed by standards for data structures, contents, value, and exchange.”

The contextualizing role of metadata supports the long-term sustainability of research data. For example, metadata serves to make heterogeneous data interoperable, such as data generated in different locations and for diverse purposes. Just as research is socially constructed – defined through the *negotiated order* of work through the relationships between individuals and technologies within the context of the *social world* (Gerson, 1983; Strauss, 1982) – so too are scientific research data. That is, data reflect the contingent values of groups and individuals, their theoretical perspectives, and their research goals. Latour & Woolgar (2013) emphasized this point, arguing that the construction of ‘facts’ is a collective endeavour that involves persuasion and argument rather than reporting a singular (objective) ‘truth.’ Measurements such as temperature and weight are highly contextual (Borgman, 2015; Tuomi, 1999). For data, context

matters because it helps us interpret data. Context also allows us to recognize data *as* data (Borgman, 2007; Edwards, 2017). For instance, a dead opossum on the side of a highway is roadkill to the average motorist. But to the opossum ecologist, the animal's remains are data for a project tracking North American marsupial migration (Walsh et al., 2017).

2.3.2 Research Data Management (RDM)

Research data management (RDM) is defined relative to the goals of the community of practice using the concept. A data preservationist defines RDM with a checklist of the data management steps needed to properly archive them (Borgman et al., 2015; Sewerin, 2015). A professional organization training researchers on data management defines RDM in terms of the research lifecycle and the expertise needed to manage data with quality and consistency (Sallans & Lake, 2014). Groups concerned with near-term project goals define RDM in terms of immediate needs for access, sharing, version control, and communication. Long-term data management may broaden the scope of the stages involved in RDM, given concern for continued use, rights management, and research replicability (Rougier et al., 2017; Sands, 2017). In this work, *data management* is defined in this broader sense and includes a range of activities within data lifecycles and "data journeys" (Bates et al., 2016): acquisition, collection, storage, processing, organization, analysis, dissemination, archiving, preservation, and reuse of data.

The conceptual and empirical work on RDM can be organized into three areas and the methods of data elicitation and analysis. The conceptualizations and major debates are grouped into the areas RDM lifecycle models, RDM workforce, and RDM expertise.

2.3.3 RDM Knowledge Infrastructures

Research data management (RDM) practices rely on supportive *infrastructures*, a term which here encompasses technological, normative, policy, and organizational components

(Scroggins & Pasquetto, 2020; Star & Ruhleder, 1996). For example, infrastructures for data management are constituted of not only the computing software and databases, but also the standards (e.g., metadata), personnel who clean the data, policies for data sharing, and norms of the scientific community about the value of data management and deposit (Borgman, 2015; Darch et al., 2020a). Infrastructure studies explain how systems that support science function and change over time, and what – and who – sustains and maintains them. Star & Ruhleder (1996) identified dimensions of infrastructures that impact on organizational change, two of which have special salience to faculty data management practices.

Disciplinary training, available resources, and institutional contexts, among other factors, shape faculty practices (Stephan, 2012a). Infrastructure studies theorize how these factors are historic, in the sense that infrastructure is not "built from scratch" (Darch et al., 2020a) but built on top of an "installed base" and inherits its advantages and drawbacks (Star & Ruhleder, 1996, p. 116). Another aspect of infrastructures is that it "links with conventions of practice," that is, an infrastructure both "shapes and is shaped by the norms of a community of practice" (Darch et al., 2020a). The workforces engaged in building and interacting with the infrastructures or the human infrastructure of cyberinfrastructure (C. P. Lee et al., 2006) are socialized and enculturated into the work context. As such, they inherit the values, cultural practices, and norms encoded into the existing infrastructures (Darch et al., 2020a). The researchers and staff, in turn, shape the following design and development of infrastructure (Shilton, 2015).

Building on infrastructure studies, scholarship conceptualized *knowledge infrastructures*. The act of producing a knowledge product or "committing to record" a scientific output such as a publication or research dataset occur in a relational web (Bowker, 2005; Edwards, 2010). These networked contexts are knowledge infrastructures, the "robust networks of people, artifacts, and

institutions that generate, share, and maintain specific knowledge about the human and natural worlds" (2010, p. 17). Data are embedded in infrastructures. They cannot not be generated outside of them because they depend on a range of entities for meaning, including policies, people, places, and technologies among other contextual features (Bowker, 2005; Bowker et al., 2016; Sands, 2017). Knowledge infrastructures for RDM exist at multiple levels and may be intended for different lengths of time. For example, a macro-level of an RDM knowledge infrastructure is the Internet whereas a local drive or remote server is another level of the infrastructure of making and managing research datasets. The varying levels and short-term and long-term needs of a data intensive scientific project can make it difficult to create RDM policies that address a pluralism of research goals (e.g., long-term preservation and meeting deadlines).

Practitioners in RDM including professional organizations, working groups, and policy reports have made recommendations to address these multiple, diverging requirements of the multi-level infrastructures for scientific data stewardship. For instance, the FAIR Guiding Principles for Research data management and Stewardship were published in 2016 to provide general goals for scientific research data. FAIR represents the goal of findable, accessible, interoperable, and reproducible scientific data (Wilkinson et al., 2016). The mission of FAIR is articulated as making data "machine-operable," prioritizing and optimizing for the automated parts of the knowledge infrastructures of RDM. Additional policy reports have been published to address the needs of professional communities that rely on RDM knowledge infrastructures, such as the Academic Research Libraries (ARL), the National Science Board, and the Joint Leadership Group of the National Digital Stewardship Alliance. The guidelines from the institutional level are part of professional efforts to address the needs of data intensive RDM infrastructures; working groups comprised of information technology professionals who support

scientific research and aim to develop best practices (e.g., Geiger et al., 2018; Workshop on Best Practices for Computational and Data Intensive Research, 2019; Exchanging Best Practices in Supporting Computational and Data-Intensive Research).

2.3.4 RDM Workflows

Data management is imprecisely defined in part because it encompasses multiple steps along a multi-faceted path. The RDM path can frequently branch into smaller tasks, making a universal definition for the step-by-step process challenging if not too broad to be useful in local contexts. Further, scientific processes can often be nonlinear, involving unexpected turns and setbacks whereby the processes for data and metadata use are revised or abandoned. Documents, data, and metadata are fluid entities that come into play as interstitial objects within an overarching praxis (Bates et al., 2016; Edwards, 2017).

To describe the stepwise albeit iterative processes of RDM, lifecycle models have been proposed by information professionals (Cox & Tam, 2018), industry practitioners (Meng, 2019), and science studies scholars (Borgman, 2019; Greenberg, 2009). The data lifecycle is a concept common within library and information sciences that acts as guide to chart the steps of data from start to an end point, whether being discarded or archived (Sands, 2017). More broadly, the research lifecycle is the series of steps following the project goals and milestones. The data management lifecycle is a subset of steps structured by these two cycles. Specifically, the data management lifecycle is a subset of the steps in the data lifecycle, which is tied to the research lifecycle. In other words, data management lifecycle models identify the data activities in the data lifecycle related to data the deployment of resources toward the goal of organizing and using data in the course of the research life cycle (Crowston & Qin, 2011; Van Tuyl et al., 2015).

Multiple life cycle models appear in the literature. They are used to depict the stages of RDM at various levels of specificity. The level granularity is contingent on the scientific discipline and RDM goals (Greenberg, 2009). Broad level models highlight the computational and conceptual activities in the research data lifecycle (e.g., Bratt et al., 2017). An example of a flowchart style of data lifecycle is shown in **Figure 1**.

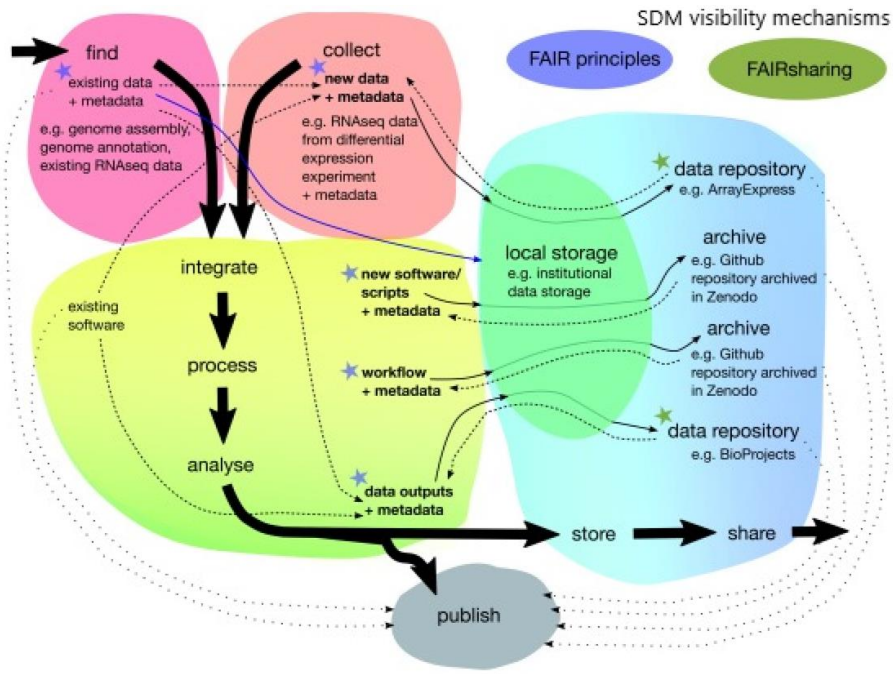

Figure 1: A Project-Level Flowchart of the Data Life Cycle (Adapted from Kowalczyk, 2018)

Although life cycle models are useful heuristic tools for understanding the major process areas of RDM, they require a discipline-specific understanding in application and are often not appropriate to be generalized. One life cycle model specific to an area of disciplinary practice is the Center for Embedded Network Sensing (CENS). This model shows detail on faculty RDM practices in the middle stages of the life cycle (**Figure 2**).

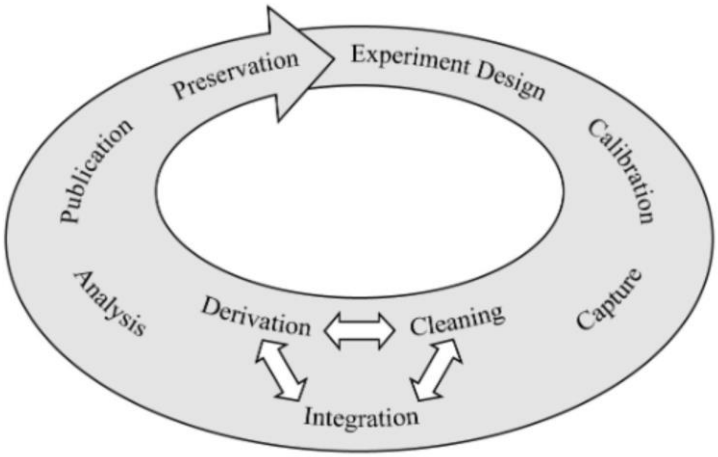

Figure 2: Center for Network Sensing (CENS) research data life cycle model (Adapted from Sands, 2017; Wallis et al., 2008)

The CENS data life cycle model (Figure 2) highlights the importance of RDM along the entire research process (Sands, 2017; Wallis et al., 2008). The data management process is continuous and data can be reused for follow-up research projects (Wallis et al., 2008).

Critics have advocated for revisions to linear lifecycle models, giving rise to models that incorporate various timescales and iterative movements between the lifecycle stages (Borgman, 2019; Cox & Tam, 2018). Life cycle models have been critiqued as reductive and linear, misrepresenting the "messy" and circuitous nature of many scientific practices and processes in RDM (Meng, 2019; Sawyer et al., 2017). Alternative models for conceptualizing the timelines of data management activities have been proposed to address these concerns. The "collaborative rhythms" of research describes multiple, overlapping and simultaneous timelines interact to shape each other as research work together (Ribes & Bowker, 2008; Sawyer et al., 2017; Steinhardt & Jackson, 2014). Relatedly, the method of 'data journeys' recently was developed to describe the temporal steps within a data intensive meteorology project. This approach is similar with the rhythms approach, though it tends also to focus on the artifacts (i.e., datasets) and their

circulation (e.g., sharing, reusing), though it highlights key personnel throughout the data journey (Bates et al., 2016).

Sands (2017) also identifies tensions between long-term planning and a dynamic and uncertain research environment as a paradox in RDM life cycle development. The turnover of personnel within the project are another part of the paradoxical tensions as a source of variability. Their roles are overlooked by lifecycle models given their focus on practices and processes, obfuscating the human infrastructure of cyberinfrastructure (C. P. Lee et al., 2006). The next section address this, turning to focus on the workforces of research data management.

2.3.5 RDM Workforces

Workforces are a vital part of the management of scientific research data. Long-term data sustainability requires allocation of resources to workforces, technology, and expertise for RDM. A purely technical solution is insufficient for research data management (M. Baker, 2016; Ray, 2013). As Paisley (1968) reminded us in the late 1960s in his letter to the Vannevar Bush's vision for the OSRD, "*As We May think, Information Systems Do Not"* (Paisley, 1968). Competencies for data management are therefore critical and workforces across the data lifecycle play central roles. As data intensive scientific disciplines continue to emerge, there will thus need to be a heightened level of "advancement of digital curation, and therefore in the digital curation workforce" (Griffin et al., 2018; Larsen et al., 2014; Sands, 2017). However, RDM workforces can often be rendered invisible by models that do not represent them when appropriate and beneficial for the goals of all RDM stakeholders (Borgman, 2015). A lack of support from RDM expert workforces threatens "the danger that data will be created in unusable forms, managed inappropriately, or stored ineffectively" (Research Information Network in Sands, 2017).

The importance of cultivating RDM expertise has been articulated by multiple stakeholders. Less clear, however, is what practices will be carried out by which professional populations. Data management expertise is not concentrated in a single person but dispersed among multiple professional roles along the value of chain of scholarship (Borgman, 2015). For example, database administrators bring expertise at the phases of data collection and information architecture. Librarians have traditionally taken responsibility for data curation and archiving, the end stages of the data life cycle. In between the beginning phase and the end phases of RDM are faculty research efforts to manage scientific data. Faculty RDM practices are often executed within the ongoing "mess" of research-in-progress (Østerlund & Carlile, 2005; Sawyer et al., 2017). While faculty are trained in their domain and specialization, RDM expertise is not a part of most formal science education (Griffin et al., 2018). Information professionals can fulfill some of the faculty RDM requirements, but tend to be trained the later stages of the cycle (Ray, 2013).

In this dissertation, *research data management workforce* is defined as the workforces responsible for "managing, stewarding, sustaining, serving, storing, archiving, curating, or preserving scientific research data" (Sands et al., 2014; Sands, 2017). The next sections review the conceptual and empirical literature in faculty data management practices, expertise, process areas, and best practices, with a focus on genetics and genomics. I identify the process areas in which faculty's core RDM responsibilities and expertise lie and discuss the expertise shift to a more computational, digital, and data centric faculty. Best practices in data intensive biosciences on genetics and genomics conclude the section.

2.3.6 RDM Faculty Practices

Faculty are a central constituency among the workforces for RDM. Research data quality relies on data management skills among all stakeholders in the distributed division of labor in

RDM. As such, faculty need to be a strong link in the RDM chain of data management to ensure sustained data quality. In the data management chain and broader research data lifecycle, faculty play a crucial role given their influence on the research process and their role in data quality assurance. Faculty tend to have a high level of organizational control over much of the data management process, given that they often set the research agenda which, in turn, influences: what data is acquired, how it is processed; what is the data quality, how to design the analysis; when and how to conduct data documentation; how to interpret the research significance of the data; select the publishing venue, and the appropriate contextual metadata and document to accompanying the archival dataset.

The range of faculty responsibilities and expertise is broadening with data intensive research. New roles for faculty emerge as team sizes grow, novel technology is introduced, and the demands of RDM become urgent given the potential for loss of data (Ray, 2013; Sawyer et al., 2017). The shift to a research agenda that require more collaborative activity, complexity in the topic, and of a data centric nature has demanded greater expertise from faculty researchers. Further, faculty RDM practices, processes, and corresponding *best* practices are discipline specific. Studies show how the local RDM practices of faculty are shaped by a hybrid of broad institutional RDM directives and local idiosyncrasies, such as faculty management style (Whitmire et al., 2015), the data culture of the lab (Thursby et al., 2018a), and the nature of a research project itself (Darch et al., 2020c). That is, what is a best practice for RDM in a small data intensive bioscience project may be vastly different from the RDM best practices for that of a large astrophysics lab (Darch et al., 2020; Sands, 2017).

Among bioscience faculty members, the conversation around best practices remains heuristic (Griffin et al., 2018; Schneider et al., 2019; Williams & Teal, 2017). The process

models for RDM best practices are high-level and do not account for discipline-specific data culture, faculty expertise and experience, and other factors unique to a faculty-led research group. In this section, the literature is reviewed in the practices of faculty, existing models for cataloguing and assessing them, and the perceptions of faculty about their own data management needs, challenges, and the opportunities for improved RDM.

Academic library surveys have been conducted to ascertain the needs of several scientific disciplines to document the data management practices and preferences of faculty (Akers & Doty, 2013; Sewerin, 2015; Van Tuyl et al., 2015; Whitmire et al., 2015). The goal of these often campus-wide surveys was to compare the disciplinary differences for RDM preferences in data management practices and services based on disciplinary data cultures and local lab practices (Akers & Doty, 2013). In multiple surveys, the faculty members surveyed represented general disciplinary areas, e.g., arts and humanities, social sciences, medical sciences, basic sciences.

The surveys inquired about multiple process areas of RDM across the RDM life cycle. For example, the survey asked faculty about Data Storage and Back-Up, a set of activities related to allocating digital storage space, the quantity of digital data stored (e.g. in gigabytes), and the method for storage (e.g., university services, lab-based storage) (Akers & Doty, 2013). Another process area measured by multiple surveys was Data Management Planning (Akers & Doty, 2013; Sewerin, 2015; Whitmire et al., 2015). The survey construed this as the level of awareness faculty have of funding agency requirements for research data management (e.g., NSF and/or NIH policies). Akers & Doty (2013) found faculty researchers across all disciplines to be only somewhat familiar or not familiar at all with requirements for RDM and data sharing plans, as related to granting agencies.

Surveys frequently asked about faculty RDM practices depositing and finding data in databanks. Authors found the most commonly used are those provided by the National Center for Bioinformatics (NCBI) (e.g., GenBank, Sequence Read Archive (SRA), Gene Expression Omnibus (GEO), dbGaP, and Protein Data Bank (PDB) (Akers & Doty, 2013; Van Tuyl et al., 2015). Documentation was a major process area that was examined in terms of faculty perceptions, competency, and variability between disciplines (Akers & Doty, 2013; Sewerin, 2015; Whitmire et al., 2015). For instance, surveys asked "how familiar are you with documenting and/or creating metadata for your data so the contents are readable by machines and others?" (Akers & Doty, 2013; Sewerin, 2015; Whitmire et al., 2015). In many of the surveys, faculty were most interested in RDM assistance in preparing the RDM plan for grant applications and workshops on data management practices (Van Tuyl et al., 2015).

As the quantity of research data increases, so too must the qualitative nature of data education adapt (Sands, 2017). In an analysis of the exponential rate of data expansion to anticipate science computing futures, Alexander Szalay and Jim Gray (2006) argue the current practices of RDM in research groups is approaching their limits. They argue "today's graduate students need formal training in areas beyond their central discipline: they need to know some data management, computational concepts and statistical techniques" (Szalay & Gray, 2006). To reconceptualize expertise in data intensive research, a metaphor of research knowledge has been invoked, describing the "I-shaped," "T-shaped," and "π-shaped" researchers (Michels, 2017; Sands, 2017; Xconomy, 2013). Highly specialized researchers with deep knowledge are I-shaped, whereas "T-shaped" have "broad and deep" knowledge (Benderly et al., 2008). As such, the "π-shaped" researcher emphasizes strengths needed from two fields, that is, knowledge in the scientific domain knowledge and technical knowledge (Michels, 2019; Szalay 2017; Sands,

2017) (**Figure 3**). This analogy models the breadth and depth of expertise in T, Π, Γ, and M form, where researcher knowledge spans the scientific, computational, and statistical domains.

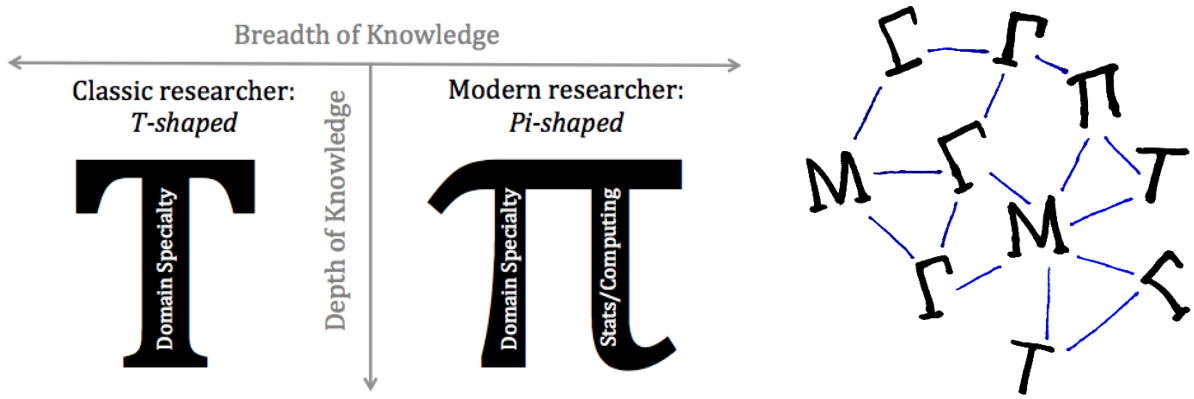

Figure 3: Research expertise of data intensive scientists. Adapted from Jake VanderPlas (2014) and Fiore-Gartland & Tanweer (2017).

Scholars and corporate leaders argue there is a lack in formal education for upskilling of scientists to gain "π-shaped" expertise (Michels, 2019; Xconomy, 2013). That is, increasingly, workers must possess both broad management abilities and deep specialized skills to become a π-shaped researcher (Benderly, 2008, 2013; Benderly et al., 2008; Sands, 2017). Critics call for a revision of graduate-level programs to generate researchers that acquire statistical and computational methods that help drive modern research. Studies of these types of researchers suggested there is a network of multiple types of expertise (Fiore-Gartland & Tanweer, 2017; Venkatraman, 2013).

In this dissertation, the *scientific domain knowledge* of interest is the data intensive biosciences, namely, genetics and genomics faculty as included in formal education for these disciplines. Within the educational program, data science expertise is often embedded as part of the methodological education of scientists in biology, given the decades-old development of statistical and data-intensive techniques in genetics research, such as DNA sequencing (Durmaz et al., 2015; Mardis, 2008). The computational skills and technical expertise, however, frequently

are found outside the disciplinary curriculum such as through self-initiated learning by the student, experience in a technical career, or pursuit of a computer science degree (Michener & Jones, 2012; Williams & Teal, 2017).

The expertise required for research data management is closely aligned with the skills and experience of so-called "π-shaped" researchers. That is, at minimum, two disciplines are necessary for faculty in data intensive science to plan for integrated data management infrastructures within their lab, and connected to the wider RDM ecosystem (Larsen et al., 2014; Ribes & Bowker, 2008). Current paths for undergraduate, postgraduate training and professional development, e.g., Research Data Management certification, either are nascent (e.g., in information science schools) or nonexistent (Demchenko & Stoy, 2021; Read et al., 2019). This is the case even in STEM sciences with relatively mature infrastructure for data management and sharing (e.g., genomics). As Griffin et al. (2018) argue: "learning how to find, store and share research data is not typically an explicit part of undergraduate or postgraduate training in the biological sciences" (Campbell, 2009; Strasser & Hampton, 2012; Tenopir et al., 2016). Subdomains in the biosciences such as ecology have established a body of professional guidance for students in data management (Alidina et al., 2008; Whitlock, 2011). Nonetheless, scientists and RDM researchers have decried the lack of formalized education for how to manage the complexity, scope, and heterogeneity of research data.

In genetics and genomics, data intensive activity across the small and big science efforts thus require data management expertise (Bala & Gupta, 2010; Durmaz et al., 2015). Faculty RDM involves dual expertise in the biosciences as well as in a digital ecosystem requiring computational skills (Darch et al., 2020c; Ribes & Finholt, 2009; Williams & Teal, 2017). However, there are trade-offs involved in becoming a researcher with depth of knowledge in the

domain and the computational expertise central to RDM. Expertise takes time to develop, and the iterativity of computational "hacking" requires time; so too does publishing and achieving traditional tenure and promotion milestones (Borgman, 2000, 2015). Because publications remain an important indicator for research evaluation, the work of RDM can often be at odds from effort allocated toward publishing (Levine, 2014; Ribes & Finholt, 2009; Star & Ruhleder, 1996). The RDM work is often delegated to technicians and data managers, who tend not to pursue tenure-track positions though their expertise and experience satisfy much of the criterion for faculty positions (Geiger et al., 2018; Shapin, 1989). Similar to astronomy, the biosciences is a domain where career paths, formal education, and professional development for RDM are developing, but not widely institutionalized (Sands, 2017). Instead, scientists find RDM expertise through collaborations or ad hoc means (Akers & Doty, 2013; Diekema et al., 2014).

Data management work is often a part of the creating and maintaining knowledge infrastructures (Edwards, 2017). Work that is seen as infrastructural "maintenance" can often be rendered "invisible" (Paisley, 1968; Shapin, 1989). Data management is often perceived as infrastructural maintenance. Therefore, faculty doing the infrastructural maintenance work of data management can be considered as *articulation work.* RDM is an integral part of data-intensive science. Yet, but fades into the background because it is taken for granted and "[becomes] invisible by virtue of routine. If one looked, one could literally see the work being done…but the taken for granted status means that it is functionally invisible" (Bowker & Star, 2000; Sands, 2017; Star & Ruhleder, 1996). Accordingly, the workforces doing the work of can be invisible, and include system administrators, faculty, long-term data managers, software engineers, and those who maintain existing infrastructures (Borgman, 2015; Bowker & Star, 1999; Ribes & Finholt, 2009; Ribes & Jackson, 2013; Star & Strauss, 1999).

Faculty RDM expertise is closely tied to the institutionalization of RDM processes and professional practices of faculty. The fourth paradigm involves an increasingly data-intensive scientific workforce that may appear to have a mature knowledge infrastructure for supporting RDM. However, there remains work to be done in building scholarly research infrastructures, process maturity and automation, and continuing to develop a community that works together to continually address emerging opportunities and challenges posed by the fourth paradigm (Griffin et al., 2018; Williams & Teal, 2017). The need to build and reinforce RDM expertise requires more than technical solutions, but also science policy and reward systems that align with long-term data management goals. Cyberinfrastructure (CI)-enabled science data intensive disciplines have indeed benefitted from the investments in technical systems, funding initiatives, and training programs for technologies and workforces in data intensive science.

2.3.7 Section Summary and Discussion

In this chapter, definitions, conceptual models, and empirical examples of research data management (RDM) were reviewed. Research data has outgrown the capacity of many existing technical infrastructures and management services. In response, scholarly communication studies have called for deeper understandings of early stages in data management life cycles by faculty scientists and related workforces. Although research data management (RDM) has been substantially improved because of a widespread concern for data sustainability, the local levels of data management remain highly site-specific and disunified. The organizational efforts to improve RDM have occurred in disciplines with community-wide needs for data standardization.

Literature largely agrees that RDM requires a workforce with expertise that has creating shifts in professional training requires from a T-shaped to a pi-shaped researcher, with greater data and computational expertise. Ample studies in LIS and organizational studies show RDM is

increasingly important which has led professionals to develop more data management guidelines, standards, and best practices to systematize RDM (Abrams et al., 2009; Darch et al., 2020c). Policy reports offer high level guidance and generalized recommendations, and working groups provide context-specific guidance that address immediate needs of a local group. Professional recommendations are important in the design of policy and technologies.

However, these tended to be idealized versions of actual researcher practices. Likewise, workflows models are high-level abstractions of the daily practices of data management. They are not empirical findings or theoretical frameworks that advance conceptual discourses in the areas of scholarly communication studies. As Sands (2017) argues, the generalities of policy reports "necessitate complementary empirical studies focused on the intricacies of scientific data practices" (p. 14). Empirical studies in this area tend to represent practitioner perspectives which focus on the end of the research lifecycle: data curation, archiving, and services to support RDM.

These studies made valuable contributions found differences between disciplines the level of systematization of data management, suggesting some reasons why RDM is systematized or more "mature," that is, institutionalized. The institutionalization of RDM has crucial implications for sustainable data preservation goals, research process transparency, and the training of emerging RDM workforces at various institutional and temporal scales and big and small sciences alike. The next section reviews perspectives and empirical studies of the institutionalization of research data, with a focus on U.S. faculty data management practices.

2.4 Institutionalization of RDM

The previous section reviewed RDM conceptualizations, process models, and library conceptualizations of RDM. These studies compare the disciplinary differences in RDM practice based on literature reviews, and largely represented the data curator perspective. In this section, I

cover the foundations of institutionalization to locate this study within the multiple strands of traditional and neo-institutional theory. . Here, the faculty perspective on components of institutionalization of RDM is brought to bear, e.g., faculty perceptions of institutional pressures on RDM. In describing the historical underpinnings, I cover the key tenants and influential ideas such as the three institutional pillars and institutional logics. I then narrow the scope to focus on the context of science, specifically, the institutionalization of research data management and deposit of RDM and data deposit. I also review empirical studies of the outcomes and impacts of institutionalization of work practices with a focus on the impacts of formal models of workflows and research data management and sharing.

2.4.1 Foundations of Institutional Theory

Institutional theory is a body of work originating in sociological studies. It is used in this study to analyze the institutional pressures and forces shaping RDM and data deposit. Institutional theory came from sociological traditions interested in explaining the behavior of organizations (W. R. Scott, 2013). Since the early 1990s, institutional theory was extended to become neo-institutional theory. Neo-institutional theory is what I use in this study. Neo-institutional theory extends beyond organizations to include, and focus on, individuals (W. R. Scott, 2013). Contemporary institutional theory researchers originally developed theory to explain how the wider organization milieu influences individuals through an imposition of pressures from the institutional context. Neo-institutional theory posits, then, that individuals face pressures to conform with to agreed-upon notions what behavior is appropriate.

Several analytical concepts are important for understanding institutions and institutionalization. The analytical concept of the *field* is crucial to understanding an institution and institutionalization (note the distinction) and is defined nicely by Scott (2013). The concept

of the *field* originates from Dimaggio and Powell (1986). Important work in this area is has been conducted by such scholars as Walter Powell, Jeanette Colyvas, and Hoyku Hwang. *Institutionalization* is the process whereby practices and/or values become taken-for-granted or legitimate. Processes of institutionalization within a field across time and space build up the regulative, normative, and cultural-cognitive pillars. This has been described in the literature as the micro-foundations of institutional theory.

Adhering to these normative expectations help researchers obtain resources like financial, human, and cultural capital through the process of achieving *organizational legitimacy* (Tolbert, 1985). *Organizational legitimacy* is defined as the alignment of an organization's actions and values with the values in the wider context of the field and society deems appropriate (Dowling & Pfeffer, 1975). That is, in making choices, individuals and "social actors" do not merely make a rational calculus based on what will be most productive or cost-saving, but also consider how their behaviors will impact their perceived legitimacy. Thus, organizational scholars consider organizational legitimacy a resource and a constraint. These analytic concepts provide a foundation for understanding institutional theory, and its progression to neo-institutional theory. I focus on three concepts central to the RQs: (1) *Institutions and institutional logic,* (2) *institutional pressures,* and (3) *institutionalization*.

2.4.1.1 Institutions and Institutional Logics

Institutions are social structures constraining actors' decisions. They are the taken-for-granted informal and formal rules about what is appropriate behavior and choices. As an analytic concept, they can identify predictable conditions to explain these choices (Greenwood et al., 2017). There are many definitions of institutions, depending on the level of organizational complexity and formality. The broadest definitions can include informal but highly regular

patterns of activity, e.g., handshakes, while the narrowest conceptualizations require there to be formal mechanisms such as laws, bureaucratic procedures, and specific rules.

Scott (2001) defined institutions as “social structures that have attained a high degree of resilience” (p. 48). Scott’s (2001) definition points to a similarity among many definitions: the importance of the stability and persistence of an institution as well as a focus on institutional change. As Scott (2001, p. 48) explains:

> [Institutions] are composed of cultural-cognitive, normative, and regulative elements that, together with associated activities and resources, provide stability and meaning to social life. Institutions are transmitted by various types of carriers, including symbolic systems, relational systems, routines, and facts. Institutions operate at different levels of jurisdiction, from the world system to localized interpersonal relationships. Institutions…connote stability but are subject to change processes, both incremental and discontinuous.

Such stability and persistence of institutions is achieved through processes of *institutionalization*, a process wherein behaviors and rules become taken-for-granted and legitimized (Meyer & Rowan, 1977; Tolbert & Zucker, 1983). Notions of legitimacy that are constructed by institutions inform and shape individual’s beliefs – both about the legitimacy of the organization, but also about how an one should conduct themself (e.g., what are ‘proper’ behaviors and attitudes).

When these beliefs are widely shared among a collective, they are termed *institutional logics* (Barley, 1986). *Institutional logics* can be defined as a set of collectively constructed assumptions, beliefs, rules, and practices that provide principles to help people interpret their surroundings and conduct themselves. Institutional logic comes from organizational studies and sociological theory. It grew in the area of marketing, directing attention to how belief systems shape people’s thoughts and actions. Thornton & Ocasio (2008) define institutional logics as: "socially constructed, historical patterns of material practices, assumptions, values, beliefs, and rules by which individuals produce and reproduce their material subsistence, organize time and space, and provide meaning to their social reality" (Thornton & Ocasio, 2008, p. 804).

Logics are material and symbolic. They shape how individual actions see themselves within an organization because logics providing collective identities for community members to draw upon. Friedland and Alford (1991) elaborated how central logics are "a set of material practices and symbolic constructions – which constitute its [organizing] principles and which are available to organizations and individuals to elaborate" (p. 248). Logics exist across multiple levels: they are enacted by individuals, but reinforced at meso- and macro-levels by rules and policies (Lounsbury, 2001). In this interplay between macro-level institutional logics and individual action, logics are ultimately enacted by individuals (Zilber, 2002).

2.4.1.2 Institutional Pressures

Pressures emerge from the institutional environment to shape behaviors. By "pressures" institutional theorists refer to expectations, norms, standard operating procedures, and taken-for-granted beliefs about what is appropriate behavior. If actors do not follow these expectations and norms, they risk losing their legitimacy (DiMaggio & Powell, 2000; Heugens & Lander, 2009). Scott (2001) identified these expectations as *pressures,* conceived as "the three pillars of institutions." The three institutional pillars are summarized in **Table 2**.

Table 2: Three Pillars of Institutions (Scott, 2013). Adapted from: Koulikoff-Souviron & Harrison (2008)

	Regulative	***Normative***	***Cultural-cognitive***
Basis of compliance	Expedience	Social obligation	Taken-for-grantedness Shared understanding
Basis of order	Regulative rules	Binding expectations	Constitutive schema
Mechanisms	Coercive	Normative	Memetic
Logic	Instrumentality	Appropriateness	Orthodoxy
Indicators	Rules, Laws, Sanctions	Certification, accreditation	Common beliefs Shared logics of action
Basis of legitimacy	Legally sanctioned	Morally governed	Comprehensible, recognizable, culturally supported

The three pillars are (1) regulative, (2) normative, and (3) cultural-cognitive. *Regulative pressures* include the coercive aspects of institutions, such as laws or rules, which regulate and constrain actors' behaviors (Scott 2001). The regulative pillar forces compliance through fear of sanctions for disobedience (Scott 2001). Regulative pressures are defined as "both formal and informal pressures exerted on organizations by other organizations upon which they are dependent" (DiMaggio et al. 1983). The regulatory pressure provides individuals with governmental or authoritative power which regulates individuals' behaviors (Scott 2007). Previous studies found that on an organizational level, regulative pressures stem from diverse sources: resource dominant organizations (e.g., suppliers), parent corporations, and regulatory bodies (e.g., government) (Teo et al. 2003). Regulative pressures are sometimes explicitly written as rules and sanctions (Scott 2001).

Normative pressures can be defined as the legitimizing means that stem from collective expectations in a particular institutional context (DiMaggio et al. 1983; Scott 2001). Scott (2001) argued that normative pressures, as collective expectations, are important mechanisms to determine appropriate and legitimate behaviors in a community. Collective expectations become shared norms through training, education, and association (DiMaggio et al. 1983). The main institutions that exert normative pressure include the research community, local networks, affiliations, and certification agencies which espouse public values (Heinrich et al. 2004). Actors are likely to adjust their behaviors according to their beliefs about what other members in the same community view as appropriate (Deephouse 1996).

Cultural-cognitive pressure is a mimetic mechanism that occurs "when an organization imitates the actions of other structurally-equivalent organizations that occupy similar economic network positions in the same industry" (Burt 1982). Cultural-cognitive pressures have two main

components: the prevalence of a practice in an industry and the perceived success of high-status organizations in an industry (Haveman 1993). Cultural-cognitive pressures push social actors to copy other successful and high-status actors' practices and behaviors because they believe those successful actors' actions are more likely to produce positive results (DiMaggio et al. 1983).

In addition to Scott's (2013) discussion of the three pillars and explanation of how they pressure on social actors, he also has a detailed discussion of *institutional levels*. These levels range from practices all the way up to the global. In Scott's words, "institutions operate at different levels of jurisdiction, from the world system to localized interpersonal relationships" (p. 248). The levels are not isolated from each other but are co-constituted of actors and institutions which each shape and influence the other. Emerging directions in neo-institutional theory attempt to trace these processes from multi-level perspectives.

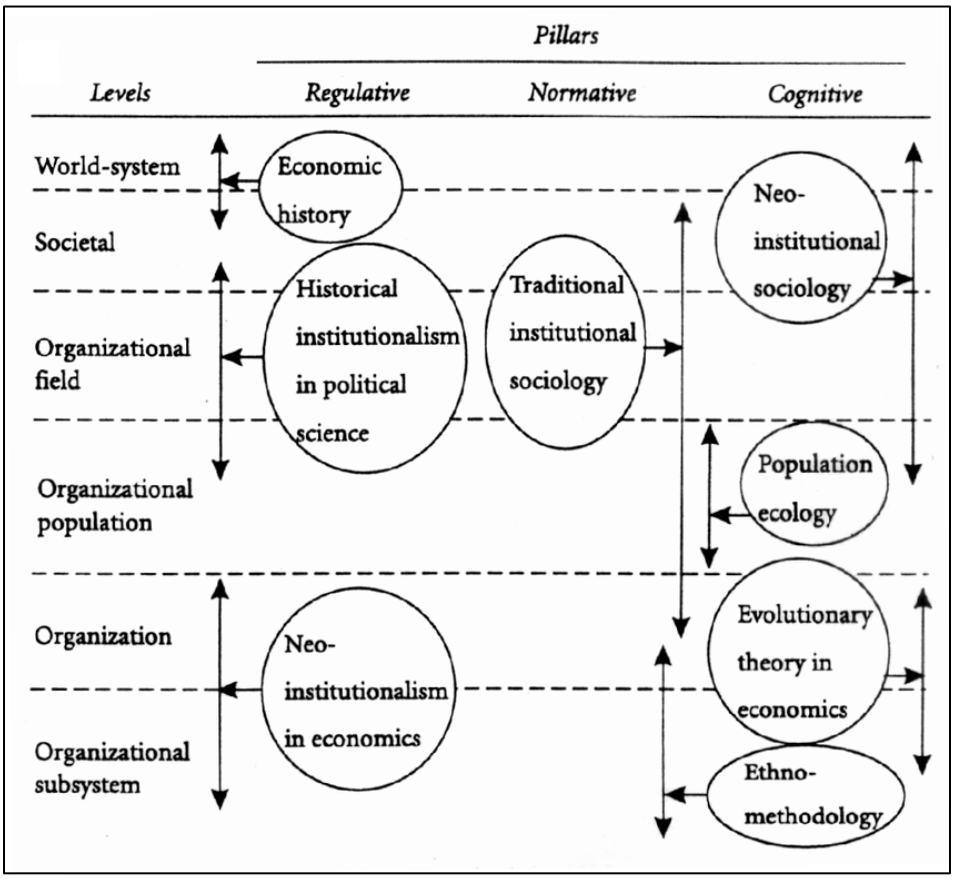

Figure 4: Institutional levels, and pillars, and fields of study

Scholars are developing new concepts to deepen the focus on neo-institutionalism on the agency of individuals, such as *institutional work*. Compared with logics, *institutional work* gives more "agentic power to social actors" and assumes that individuals can exert power in organizations (Lawrence & Suddaby, 2006). Actors can maintain, disrupt, or transform institutions through institutional work.

2.4.1.3 Institutionalization

Institutionalization is a process in which values, practices, social roles, and or concepts become legitimate (Keman, 2017). The pillars of institutions (regulative, normative, and cultural-cognitive) are built up over time and space by the processes of institutionalization within a field. This accumulation of institutionalization to form the pillars has been described in the literature as the *micro-foundations* of institutional theory (Scott, 2001). Institutionalization regulates behavior at a supra-individual level, that is, at the level of society and within organizations. The process of institutionalization can be distinguished by the presence of three steps: installing or constructing rules; developing best practices, and replacing existing rules with new rules (Keman, 2017).

By the early 20th century, the German sociologist Max Weber had theorized institutionalization and its variations. According to Keman (2017), Weber theorized a "dual logic institutionalization" by making a distinction between different types of rule configurations in the process of institutionalization: goal-oriented and idea-consolidating. According to Weber, the goal-oriented component of institutionalization is one that aims to achieve particular ends and idea-consolidating is one representing values or embodying a value-system. Keman (2017) gives the example of the division of powers to illustrate this point, where the idea of three branches of government is "both an organizational framework that results from and influences competitions of political actors and as an attempt to safeguard a certain conception of liberty" (p. 2).

Thus, institutionalization is an activity that establishes, maintains, adapts, and transforms rules and behaviors. In doing so, it impacts how individuals, organizations, and polities interact. A useful illustration of the interaction between these entities and the dynamics of *goal-oriented* and *idea-consolidating* rule configurations institutionalization is liberal democracy. Establishing and evolving a liberal democracy is not a static or one-time event, but an ongoing institutionalization process. In terms of idea-consolidation, it embodies the shared ideas that people have as to the right to civil and political protections under the law. In terms of goal-orientation, it manages the relationship between the state and organizations/individuals by creating laws and tenets to define how governance will occur.

Legitimation is also a concept closely tied to institutionalization. In fact, according to Greenwood et al. (2017), empirical studies in organizational sociology and social psychology have developed models of legitimation which parallel those of institutionalization, suggesting a more general process. Both legitimation and institutionalization share stages of installing rules, adapting them, and then going through change. For example, in their examination of the institutionalization stages over time, Lawrence et al. (2001) created 4-stages model of legitimation constituted of: "innovation, local validation, diffusion, and general validation."

2.4.2 Institutional Context of RDM

Research data management (RDM) and sharing have undergone a process of becoming increasingly embedded within the U.S. research environment, e.g., data sharing mandates and the open science movement. To provide a background of the extent and impacts of institutions, pressures, logics, and institutionalization of RDM and data deposit on the organization of U.S. faculty research data practices, I focus on the applications of (neo)institutional theory to academic research data management. The sub-sections covered to do this are *institutional*

pressures of RDM and *institutional logics* of RDM, including the norms and counter-norms of RDM and empirical studies of the institutionalization of RDM.

2.4.2.1 Institutional Logics of RDM

Neo-institutional theorists posit that institutional logic shapes researchers' beliefs, decisions, and behaviors by structuring incentives (Luo, 2007). As Paula Stephan has argued in her book *How Economics Shapes Science*, researchers are motivated by the available 'carrots and sticks' that structure the decision-space in research institutions (e.g., universities) (Stephan, 2012a). Institutional logics can assist to identify the prevailing norms, rules, and what is considered rational and appropriate conduct in data management.

Of this empirical neo-institutionalist research, one stream focuses on how logics are drawn from as sensemaking devices by central decision-makers, resulting in "logic-consistent decisions" on a specific problems and solutions (Thornton et al., 2005). There is also a body of work that has focused on the prevailing logics and incremental and more abrupt changes from one logic to another (e.g., Lounsbury, 2001; Suddaby & Greenwood, 2005). An exemplary study showcasing this is Scott et al. (2000), who showed how changes to the logics of healthcare led to "the valorization of different actors, behaviors, and governance structures" (Scott, 2001, p. 243). Thornton et al. (2005) focused on a shift in the logics of publishing in U.S. higher education to detail how the shift from professional logics to market logics led to correlate changes the ways leadership replacement was conducted (e.g., Ocasio, 1994).

Although there are few published works explicitly concerned with institutional logics in ASIS&T and *Science and Technology Studies* (STS) journals (*4S, ST&HV*), a few do focus on the values and logics underlying contemporary data management, sharing, and long-term sustainability. The first is in relation to the values and logics of the *open science* movement (e.g.,

Mirowski, 2018). The second is concerned with the institutional logics of research software sustainability (e.g., Weber, 2020). These two empirical areas of study shed light on the institutional logics that are implicitly and explicitly contained in the values that are constitute their logic(s). In other words, understanding the institutional factors that shape data management in U.S. faculty data practices relies on understanding the values and norms of science. Values and norms create rules for what is rational and appropriate to justify practices and policies.

In his book, *Priorities in Scientific Discovery,* Merton (1957), described the priorities and values which academic researchers hold in their work as scientists and scholars, similar toto Paula Stephan's book *How Economics Shapes Science.* (Stephan, 2012a). Both of their research finds that scholars place high value on being "the first" to create or discover a scientific issue of community importance (Stephan, 2012a). Second, scholarly reputation is a value of high importance for scientists, often quantified by the number of publications and citations for those publications (Merton, 1957). Motivations are indirect indicators of what researchers' value. Motivations also include solving complex problems or "puzzles" (Stephan, 2012a) as well as mentoring students. The reward system for science, such as crediting and compensation (e.g., via tenure and promotion), also reflect and embodies these scientific values, priorities, and motivations. For example, the creation of new knowledge is a value embodied in the reward system of published work (e.g., the impact of journal articles reporting research findings).

These values are socialized into new researchers when they enter the research context, constituted of the professional myths, values, and reward systems (Mirowski, 2018). Researchers then internalize the norms of science as institutionalized values (Merton 1973). For example, the value of reputation, priority, and credit for being the first aligns with the Mertonian norms, often represented by the mnemonic 'CUDOS': Communalism, Universalism, Disinterestedness, and

Organized skepticism. However, scientific counter-norms were proposed by (Mitroff, 1974) who studied the behaviors and practices of scientists and found their activities to reflect nearly opposite norms: Solitariness, Particularism, Interestedness, and Organized Dogmatism.

Table 3: Definitions and Summary of Mertonian Norms of Science (1973) and Mitroff's (1974) Counter Norms

Definitions	Norms of Science	Counter Norms	Definitions
Scientific findings must be shared with all members of the scientific community	Communalism	Solitariness	Scientists consider research findings as protected property and secrecy is needed to protect them
Scientific research must be judged by scientific criteria rather than by scientists	Universalism	Particularism	Judging scientific findings according to scientists' social backgrounds
The preference for the advancement of knowledge as opposed to the individual motives of the scientist	Disinterested-ness	Interestedness	Scientists care more about financial benefits from research than personal satisfaction and reputation
Scientific findings should be examined for empirical evidence of scientific merit before being accepted	Organized Skepticism	Organized Dogmatism	Scientists accept certain scientific findings without examining them carefully

These norms can be applied to research data management (RDM) and data sharing. As Kim (2013) summarized, researchers value communalism by distributing their research in an unlimited manner. They demonstrate universalism through beliefs that rewards of science (e.g., reputation and credit) are premised on research quality rather than pedigree. They show disinterestedness through professing the value of advancing instead of via financial compensation. The extent the community conducts peer review reflects organized skepticism.

In the open science case, these values are such as that which Merton famously called the norms of "the Republic of Science" (Merton, 1973). These norms of science are communalism, universalism, disinterestedness, originality, and skepticism. Open science has values reflecting

these norms, but also conflicts in those values. Researchers conform to these norms of science as institutionalized within the reward system to enact their values (e.g., achieving a good reputation, being the first to publish). Prior studies on data sharing indicate similar "ambiguous and conflicting logics" (Thornton et al., 2005), where Mertonian norms and Mitroff's counter norms coexist (S. Kowalczyk & Shankar, 2011; Louis et al., 2002). In genetics for instance, studies show that geneticists follow the norms of Solitariness and Interestedness (Ceci, 1988). Yet, McCain (1991) found that genetics researchers practices reflect disinterestedness and communalism. The situation of ambiguous and/or inconsistent logics one that may be a product of new organizational members or the "layering (or 'sedimentation') of new organizational imprints upon old ones over time" (Marquis & Tilcsik, 2013).

In a study of the institutional logics of sustainable software found that the current logic of long-term research software sustainability is one of a "finite game," designed so players compete to achieve an end goal. Examples of these games are traditional sports such as football or basketball. This is how research grants often work – there is a finite amount of time (e.g., 1-3 years) in which the software is developed, and an end goal in mind. However, infinite games are designed so the "state of play" is maintained, and the goal is to keep the game going (Weber, 2020, p. 202). This is a more collaborative approach, because it does not have an end and requires people to coordinate solutions to maintain game play. Weber (2020) argues that the logic needs to change to ensure scientific products like datasets and software can be sustainable over the long-term. RDM logics in genomics and other fields are premised on the disciplines from which they are embedded. The studies of norms and counter-norms provide a place to begin to unravel and identify the logics that inform practices and practice dynamics in RDM.

2.4.2.2 Institutional Pressures of RDM

The turn of the 21st century and the rise of CI-enabled science put pressure on scientists to manage and share data. Federal agencies and the rise in movements with open science goals, among other actors, pushed researchers to make their data more sustainable over the long-term including more sharable and findable, accessible, interoperable, and reproducible (FAIR) (Wilkinson et al., 2016). These pressures and incentives structured the expectations, norms, and rules for how researchers were to manage and share their data. Regulations as to how and when data should be managed and shared thus shifted the locus of control over data work from largely the researcher managing a project to a distributed network of institutional actors.

Studies in LIS have attested to this rise of pressure from the regulative side to manage data according to specific guidelines and mandates, focusing on whether faculty researchers have the expertise to write a data archiving plan (e.g., Van Tuyl et al., 2015), survey the disciplines most impacted by data sharing mandates (e.g., Akers & Doty, 2013), and the nature of the work to make their data ready for deposit (e.g., Curty et al., 2013; Sallans & Lake, 2014).

Kim (2013) examined the individual and institutional factors influencing data sharing in STEM and social science research, positing three sources of institutional pressures on scientists to share data: disciplinary associations, funding agencies, and journal publishers (**Figure 5**).

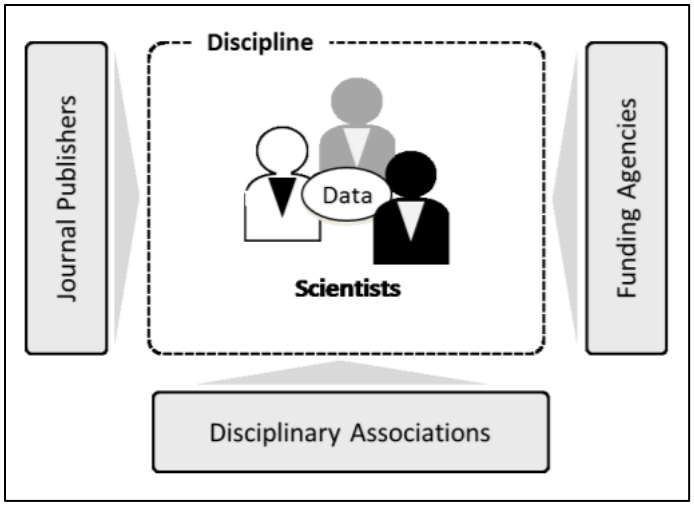

Figure 5: Scientists within institutional contexts and pressures. Adapted from: Kim (2013).

The decisions and attitudes of researchers around data management and sharing is shaped by these disciplinary culture and context, and lead to wide variation between them among scientific disciplines (Borgman, 2007; Tenopir et al., 2011). More broadly, science is considered as an institution which is both autonomous from and dependent upon other institutions. For example, researchers have the right to set their own research agendas, control their research activities free from outside strictures, and independently evaluate findings (Barber, 1952; Goldsmith, 1967; Merton, 1973; Polanyi, 1945). The value of the autonomy of science and the evidence showing the relative freedom from private or government constraints are compelling. However, it is reductive to conclude researchers and the scientific institutions are isolated from the wider social institutions in which they are embedded. As *Science and Technology Studies* (STS) researchers have emphasized, science and researchers are inextricable from society (Felt et al., 2017). Science is thus influenced by external forces such as the economy, social mores, and culture (Bloor, 1984; Cole, 2004; Shapin, 1989).

Modernization and institutionalization are closely connected. The industrial revolution impacted science as a social institution, changing the way it is structured. For instance, in post-WWII research, government funding became a main source of support for academic research.

The desire to maintain wartime speed of knowledge production and innovation led to the reorganization of science into hierarchical groups led by a project investigator (PI). Government, funding agencies and private companies are institutions which have influenced the organization and conduct of science (McGrath, 2002; Stephan, 2012a). Both the values of science and the wider institutional milieu are constituted by institutional logics which shape attitudes, behaviors, and practices related to RDM and data sharing.

2.4.2.3 Institutionalization of RDM

Institutionalization is the process whereby practices and/or values become taken-for-granted or legitimate and build up the regulative, normative, and cultural-cognitive institutional pillars (Scott, 2013). Here, *institutionalization of data management* refers to the processes whereby data practices or values become taken-for-granted and legitimized. They manifest as data policy, mandates, processes, open science ideals, and norms.

Process improvement models and best practices are examples of ways in which RDM is institutionalized. Therefore, this section draws from literature in into two areas: (1) the *practices and process areas* of RDM involving faculty; and (2) the *best practices* for faculty RDM.

The practices and process areas in which faculty are research data management are modeled include specific practices and high-level process areas. Crowston & Qin (2011) identified a high number of "key practices for RDM" which they grouped into four process areas based on the goal the practice was aimed at achieving (p. 2). To develop the key practices, Crowston & Qin (2011) reviewed literature in data science, data curation, and data management and identified twenty-one practices from data acquisition to long-term preservation. The model proposed by Crowston & Qin (2011) focuses on the practices catalogued to enable an assessment according to the capability maturity model (CMM) (Paulk et al., 1993) (**Table 4**).

Table 4: Key practices and process areas for RDM. Adapted from (Crowston & Qin, 2011)

Key Process Area	Practice	Example
Data acquisition, processing and quality assurance Goal: Reliably capture and describe scientific data in a way that facilitates preservation and reuse.	1.1 Capture/acquire data	Project data is received through a data download or transferred using storage devices such as magnetic or optical media.
	1.2 Process and prepare data for storage, analysis and distribution	
	1.3 Assure data quality (e.g., validate and audit data).	
Data description and representation Goal: Create quality metadata for data discovery, preservation, and provenance functions.	2.1 Develop and apply metadata specifications and schemas	Enter metadata for a dataset using a form-based editor. Validate data content, the adequacy of documentation, and the level to which archiving standards are adhered. Submit data, software, and accompanying documentation to designated archive. Preserve contextual metadata of for the dataset that establishes authorship and when the data was created.
	2.2 Contextualize, describe and document data	
	2.3 Document data, software, sensors and mission	
	2.4 Create descriptive and semantic metadata for datasets	
	2.5 Design mechanisms to link datasets with publications	
	2.6 Ensure interoperability with data and metadata standards	
	2.7 Ensure compliance to data standards	
Data dissemination Goal: Design and implement interfaces for users to obtain and interact with data	3.1 Identify and manage data products	Develop methods and tools for data integration and sharing with the community of practice. Plan for distribution of data products and their validation, packaging, and channels of dissemination. Identify data types and formats of data products.
	3.2 Encourage sharing	
	3.3 Distribute data	
	3.4 Provide access (e.g., by designing and piloting service models)	
Repository services / preservation Goal: Preserve collected data for long-term use	4.1 Store, backup, and secure data	Backing up databases, preserving datasets and enforcing the security of data systems Contribute to the design and community of dialogue of domain-specific databases Migrate data from previous file formats to update old records to sustainable formats.
	4.2 Manage schedules for archive generation, validation, and delivery	
	4.3 Curate data	
	4.4 Perform data migration	
	4.5 Build digital preservation network	
	4.6 Validate data archives	
	4.7 Package and deliver archives	

The process areas are clustered by the broader goals of each practice and the examples reflect the faculty roles within them. The levels of maturity of data management (DM) within a scientific research project range from Level 1 ("Initial") to Level 5 ("Optimizing") (**Figure 6**).

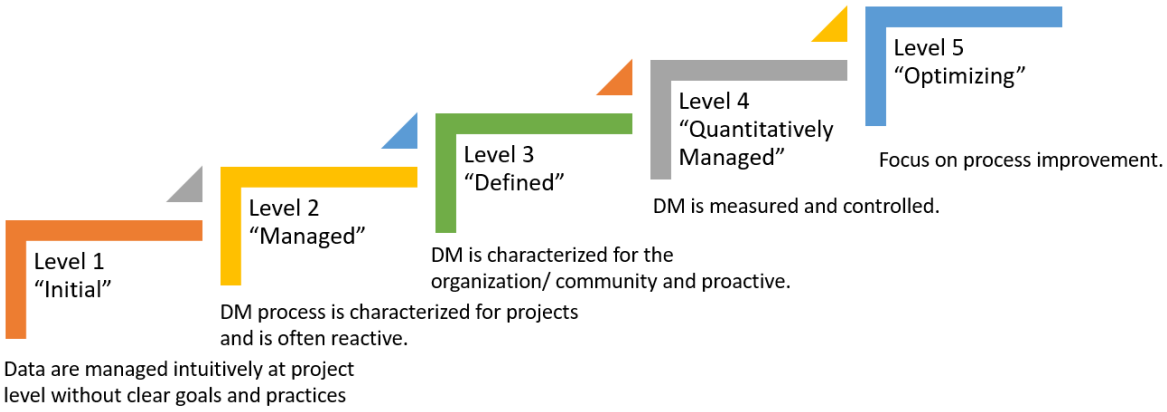

Figure 6: Capability Maturity Levels for RDM. Adapted from Crowston & Qin, 2011)

. The practices and process areas of faculty RDM literature includes organizational studies and scholarly communication literature. These tend to reflect the idealized RDM practitioner view of what faculty's roles and activities in RDM. The capability maturity model suggests a series of best practices for research data management. *Best practices* are defined as "procedures that have been shown by research and through experience to produce optimal results and that is established or proposed as a standard suitable for widespread adoption."[3] Best practices also go by *evidence-based practice* (EBP), good practices and standard operating procedures (Bardach & Patashnik, 2019).

In data-intensive biosciences, there are several voluntary guidelines which can be considered as the basis for best practices, e.g., the Bermuda Principles (F. S. Collins et al., 2003; Jones et al., 2018; Marshall, 2001), FAIR data stewardship guidelines (Wilkinson et al., 2016),

[3] https://www.merriam-webster.com/dictionary/bestpractice

and the contributor role taxonomy (Allen et al., 2014, 2019; Larivière et al., 2016). The biology community generated best practices for handling digital data in heterogeneous format (**Table 5**).

Table 5: Recommendations for bioscience RDM best practices. Adapted from Griffin et al. (2018)

Best Practices for Biosciences Research data management
• 1. Researchers reusing any data should openly acknowledge this fact and fully cite the dataset, using unique identifiers.
• 2. Researchers should endeavour to improve their own data management practices in line with best practice in their subdomain – even incremental improvement is better than none!
• 3. Researchers should provide feedback to their institution, data repositories and bodies responsible for community resources (data standards, controlled vocabularies etc.) where they identify roadblocks to good data management.
• 4. Senior scientists should lead by example and ensure all the data generated by their laboratories is well-managed, fully annotated with the appropriate metadata and made publicly available in an appropriate repository.
• 5. The importance of data management and benefits of data reuse should be taught at the undergraduate and postgraduate levels. Computational biology and bioinformatics courses…should include material about data repositories, data and metadata standards, data discovery and access strategies. Material should be domain-specific enough for students to attain learning outcomes directly relevant to their research field.
• 6. Funding bodies are already taking a lead role in this area by requiring the incorporation of a data management plan into grant applications. A next step would be for a formal check, at the end of the grant period, that this plan has been adhered to and data is available in an appropriate format for reuses.
• 7. Funding bodies and research institutions should judge quality dataset generation as a valued metric when evaluating grant or promotion applications.
• 8. Similarly, leadership and participation in community efforts in data and metadata standards, and open software and workflow development should be recognized as academic outputs.
• 9. Data repositories should ensure that the data deposition and third-party annotation processes are as FAIR and painless as possible to the naive researcher, without the need for extensive bioinformatics support.
• 10. Journals should require editors and reviewers to check manuscripts to ensure that all data, including research software code and samples where appropriate, have been made publicly available in an appropriate repository, and that methods have been described in enough detail to allow re-use and meaningful reanalysis.

As Griffin et. al (2018) state, the challenges for biology come from simultaneous, dual forces: (1) the “transition toward biology as a data science” and (2) biology’s move toward a life

cycle view of research data (p. 1). In response, best practices for RDM were developed by researchers in conferences and workshops, such as those summarized by Griffin (2018) as the steps the research community can take in ten points

The literatures reviewed here tend to be framed as idealized RDM practices. In doing so, they focus on general principles for RDM rather than the challenges and the specificities of faculty praxis. The capability maturity model for RDM by Crowston & Qin (2011) was driven by the goals of RDM and influenced by data curation practitioner ideals. The biosciences best practices literature identified the areas where RDM challenges are appearing and proposed solutions to them in 10 guiding principles (Griffin et al., 2018) (Table 5). The academic library survey literature asked faculty about RDM practices, focusing on services. Though academic library services survey studies provided a less idealized view, the survey limited RDM to the pre-defined set of survey questions delineating RDM activities (Akers & Doty, 2013; Diekema et al., 2014; Sewerin, 2015; Van Tuyl et al., 2015; Whitmire et al., 2015).

2.4.3 Section Summary and Discussion

Many scholars across multiple disciples have studied research data sharing, of how frequently researchers were sharing, their attitudes, and whether they felt supported by available services and products for RDM. Yet, the institutional context was not explicitly focused on, though it is critical for analyzing data practices. The literature is limited by seeing RDM as an isolated set of practices removed from rather than embedded in the faculty research process.

Studies highlighting the gaps in literature at large with respect to best practices for RDM faculty. The studies articulated gaps in formal training for existing workforces to improve RDM expertise. Griffin et al. (2018) argue it is unclear whose takes the role of reviewing data products, where review and validation tasks tend to be ad hoc within a laboratory or to a voluntary

workforce at smaller repositories where no centralized funding for review exists. As a maturity level for RDM, the only generic processes institutionalized were around data curation (Crowston & Qin, 2011). For all the accolades of RDM process modeling, maturity, and control, it remains to be seen what the effects are. The absence of process control may invite innovative opportunities, but when it comes to long-term sustainability of data products, mature RDM processes prove effective compared to unstructured "Level 1" processes.

There are many avenues within institutional theory that can be taken. I focus on empirical and theoretical studies of the institutionalization of RDM. I direct attention to the institutions which exert pressures to manage and deposit data such as funding agencies, journal publishers, and research data repositories. As empirical studies of institutions, pressures, and logics demonstrate, the (neo)-institutional perspective offers insight into organizational processes.

2.5 Chapter Summary

This chapter reviewed the canonical and contemporary literature in digital scholarship, research data management, and the institutionalization of RDM and data deposit. With a focus on studies of institutionalization of data work in U.S. faculty research, the chapter reviewed the relatively emergent field in a rich research tradition that has explored the institutionalization of data work in the context of U.S. academic research, and highlighted gaps and areas for further study. To address these gaps, the next chapters a 3-study qualitative study of the institutionalization of RDM and data deposit practices of faculty in U.S. research institutions. The next chapter describe the overarching research design of the dissertation, with a focus on methods, sample population, and background context shared by the three studies.

CHAPTER 3
OVERVIEW OF THE RESEARCH DESIGN

3.1 Introduction

In this chapter I outline the dissertation's overarching research methodology of a qualitative research design. The overall dissertation research design is a 3-paper study. Here, I explain the study design rationale, origins of the study, research setting, and methodological aspects shared by all studies such as the general principles of methods (e.g., interviews and grounded theory), the treatment of the data (e.g., confidentiality and data security), research contexts, and why I selected the disciplines for the population sample.

3.2 Research Design Rationale

The overarching research design of this dissertation is a qualitative research design. The qualitative approach is valuable because it enables us to ask questions about the perceptions, attitudes, norms, and dynamics of data institutionalization and articulation work. In other words, qualitative approaches allow to ask "why" and "how" questions regarding the data practices of faculty as they manage their research data within an increasingly institutionalized RDM context (Crowston & Qin, 2011; Diekema et al., 2014).

As research data management (RDM) becomes increasingly important and more institutions start RDM projects and initiatives, organizations will increasingly demand guidelines for research data management. While we cannot expect for all fields to have the same requirements for RDM, genomics can be a source of information for developing shared principles and best practices for RDM but also, more broadly, for implementing an institutionalized data process. There are fields who will need to develop RDM guidelines but who will be new to RDM concepts, and to the idea of a controlled, audited, set of standards and

evaluation criteria for their data work. They are used to flexible, intuitive, locally determined and ad hoc ways of managing their research data. The working conditions of increased guidelines for data management may be a challenging organizational transition. These fields can learn from the genomics case, a field that successfully developed, implemented, and normalized RDM policies.

Given this need, the sample populations I choose were research-active faculty in U.S. academic institutions who had submitted data to a repository from genomics, biochemistry and biophysics, and social and political science researchers. I choose the genomics community because genomics is a big science discipline with mature cyberinfrastructure for research data management (RDM) and deposit. They are a prominent example for institutional policy and scientific practice. They are a representative case and historical example of how shared principles and processes for data management and deposit developed became acceptable and imbricated within research groups. Biochemistry and biophysics are disciplines that represent the broader context of biological work but is still a related discipline. Because social scientists are not as strictly required to deposit data, they not in as rigid an incentive structure as the genomics. I leverage this difference in context to set up a comparison.

3.3 Study Origins & Motivation

This study originated from broader projects on scientific collaboration networks and studies of science in practice at the Syracuse University iSchool Metadata Lab. We used bibliometric and data science methods to analyze scientific collaboration networks and computed the statistical properties of GenBank metadata (Costa et al., 2016; Crowston & Qin, 2011; Hemsley et al., 2020). Findings from this work suggested scientists were starting to behave as if the datasets they submitted to the repository were “intellectual contributions… i.e., laying intellectual claim to their production” (Costa et al., 2016). We also found data authors were

becoming more prevalent on publications, indicating a shift among data work to a more formally recognized activity where authorship represented the intellectual contribution of the data authors.

We had a rich picture of the macroscopic and mesoscopic properties of the scale and structure of the data collaboration network (using a bibliometric methods and research policy lens to interpret the findings). However, we did not have the tools to address why data authors were contributing data to the repositories in such high numbers, what had led to the shift from no authorship metadata to a structured set of metadata. It was difficult to address questions of organizational processes and values, attitudes, and data practices associated with this new social form of a 'data author' because our project primarily uses quantitative approaches.

To address this gap, I designed a qualitative research study to pursue research questions about the values, data culture, and workflows antecedent to data deposit. The study explored how repositories shaped faculty data practices and inquire about whether and to what extent data work was being institutionalized. The qualitative study contextualized the quantitative findings.

3.4 Research Setting – U.S. Academic Research

The research setting, broadly, is U.S. academic research groups/labs. Academic research is a unique setting, distinct from, e.g., National Research Centers, because it is shaped by regulative, normative, and cultural contours that construct decision space in which faculty decide how to manage and share/deposit data (Stephan, 2012a). For example, academic research is often federally funded, a system with specific temporal rhythms (e.g., 3-year grants), policies (e.g., data sharing mandates), values (e.g., open science), and constraints (e.g., limited funds). Academic research is also shaped by the institution in which it is conducted (e.g., private college, state university), which are regulated by the state in which they are located. Faculty operate within the expected service, teaching, and research commitments of that institution. In addition,

the motivations and incentives for work in academic research are unique in that research products tend to be "public goods" (e.g., knowledge, information), which are free to all (non-excludable) and available to all regardless of how many people have access (non-rivalrous) (Stephan, 2012a). As such, the incentives to produce knowledge in academic are not usually monetary (as with private goods, e.g., an iPhone), and so other motivations come into play, such as scholarly reputation, the satisfaction of "puzzle-solving," and the prestige of being the first to say or "discover" something (*ibid*). These distinct features that make up the setting of academic research impacts how data is dealt with and how and whether it is deposited.

The sample population of interest in this setting is U.S. research-active faculty in genomics and the social sciences. Faculty researchers are key actors in the life cycles, rhythms, and journeys of data – as such, they play central roles in research data management. To set the context of the roles and responsibilities of U.S. academic faculty, it is important to understand some of the historical context of the profession. Since WWII, the job functions of an American faculty in a STEM discipline have changed substantially. The primary reason for this change can be attributed to the invention of the Principal Investigator (PI). The PI is a distinctive social form that emerged from the postwar era, a change still reflected in the operations and underlying model on which the NSF currently operates. The postwar reorganization of science implied that faculty no longer worked as solitary scientists, a sovereign under whose control the research process operates. Instead, faculty began to play a more managerial role.

The PIs of modern, data intensive research coordinate a burgeoning set of responsibilities. These roles include managing interdisciplinary collaborations among often geographically remote co-workers, given that the increased complexity of scientific problems have led a higher level of interdisciplinarity (Bärmark & Wallén, 1980; Bozeman & Boardman,

2014b; Hall et al., 2018). As a result, faculty researchers are often PIs who lead diversified, multi-skilled teams of students-in-training, staff scientists, technicians, and work with multiple collaborators (Scroggins & Pasquetto, 2020). They coordinate work and bring individual team members into alignment with institutional pressures and the demands of a data intensive research agenda, from the demands of grant funding timelines to institutional review board (IRB) directives and publishing requirements (Geiger et al., 2018). The research data collected and generated these projects are "often subject to the control and regulation of different data policies and compliance" (Crowston & Qin, 2011).

3.5 Genomics Research Data Management

Genomics is an excellent research site for studying the extent to which data deposit is institutionalized and the impact of institutionalization of RDM and deposit on long-term research data sustainability. It is a field that boasts mature infrastructures data RDM and data deposit and can inform other fields who have begun to develop guidelines for RDM but struggle with developing and implementing them in practice. As research data management (RDM) becomes increasingly important and more institutions start RDM projects and initiatives, organizations will increasingly demand guidelines for research data management. Fields who are new to implementing these guidelines, cyberinfrastructure, and policy can learn from the lessons and guiding principles of the genomics case.

Genetics research began as a 'small science,' traced as far back as the classical era with Aristotle and Pythagoras to heredity (between the 8th and 6th centuries BC) (Durmaz et al., 2015). Modern genetics is frequently traced to the experiments of Gregor Mendel on the inheritance of genetic traits in pea plants (in approximately1866). Mendel's work led to the 20th century developments, where Mendelian approaches were applied to multiple organisms, presently

referred to as 'model organisms,' e.g., the fruit fly Drosophila melanogaster (Durmaz et al., 2015). Prior to Mendel, genetics research was predominantly theoretical. Building on the advances in the domain knowledge and techniques developed in the early 1900s, genetics in the mid-20th century began an acceleration, as molecular principles were discovered.

The rise of technologies for computing, information organization, storage, and dissemination contributed substantially to genetics research, enabling landmark milestones including the Human Genome Project. The rapid technological developments in computing and communication led to what has been termed data intensive science. Data intensive genetics and genomics research data are heterogeneous, spanning a variety of molecular mechanisms, data types and formats, and a diversity of sources (e.g., environmental samples, organism genomes, and synthetic substances). Here, a focus is placed on the genetics and genomics communities engaging with databases largely devoted to curating nucleic acid sequence data because of the high level of use and the lack of focus on faculty RDM practices within data curation life cycles. Data intensive biology encompasses an array of sub-discipline, specializations, and methodological approaches. The National Institutes of Health (NIH) refer to the "biomedical enterprise" as a general umbrella area under which biological research leveraging biological data for medical and research purposes is classified (Hesse et al., 2011).

3.5.1 Genomics Data

Sequence data such as the DNA and RNA sequences are simple relative to other data types because they are represented in a machine-readable, structured text format. However, a PDF of *Moby Dick* also might be considered a form of data that is represented in a machine-readable, structured text format. What makes genomics distinct for reuse is that scientists easily interpretable by both humans and machines to address scientific questions. For instance, DNA

sequences are represented as four letters in nucleic acids base pairs: adenine (A), cytosine (C), guanine (G), and thymine (T). Scientists "know the code" of the base pairs, that is, their biological significance, so the implications for scientific questions are clearer. There are also clear data reuse cases. For example, SARS CoV-2 data was reused during the COVID-19 pandemic to conduct comparative genomic analysis (Cyranoski, 2021). As well, genomics data has clear reuse cases such as identifying functional elements of proteins, e.g. in proteomics (Hamid et al., 2009; Merrill et al., 2006) and by using model organisms (Leonelli, 2010; S. G. Oliver et al., 2016).

Because the data are represented as finite string in predictable patterns of characters, the design of the metadata annotation for some types of genomics data, e.g., model organisms, was relatively more straightforward than, e.g., multimedia (Bala & Gupta, 2010; Benson et al., 2017; Nadim, 2016). However, as Bietz & Lee, 2009 point out, the question of *what* metadata to capture is complicated by the need of re-users for an unspecified range of information about the context of collection, e.g., environmental factors. Metadata development is difficult, then, because it is hard to anticipate what a data re-user will need (Bietz & Lee, 2009). Smaller scale, rare, and heterogenous genetics data in particular suffer this conundrum (Arias et al., 2015).

3.5.2 Genomics Data Repositories

Within data intensive biology and biomedicine, the genetics and genomics communities have cohered as an active collaborative network around research data repositories. A *genetic database* has been defined as a single or a set of collective services (e.g., metadata cataloguing), products (e.g., software), and artifacts (e.g., datasets) structured to enable the search and discovery of "genes, gene products, variants, phenotypes… to enable users to retrieve genetic data, add genetic data and extract information from the data" (Durmaz et al., 2015). Data-sharing

through genetic databases not only promotes reproducibility, but also creates opportunities for novel discoveries through dataset combination and reuse (Leray et al., 2019).

Genomics went through a "communication regime change" in data deposit, where the responsibility to deposit data in a database such as GenBank shifted from the repository staff to the scientists themselves (Ankeny & Leonelli, 2015). Prior to the 1990s, librarians at data repositories such as GenBank would curate genetic datasets from the biological literature. This was referred to as 'abstracting' because staff would find datasets by reading the abstracts of publications to identify datasets to cull and annotate (**Figure 7**).

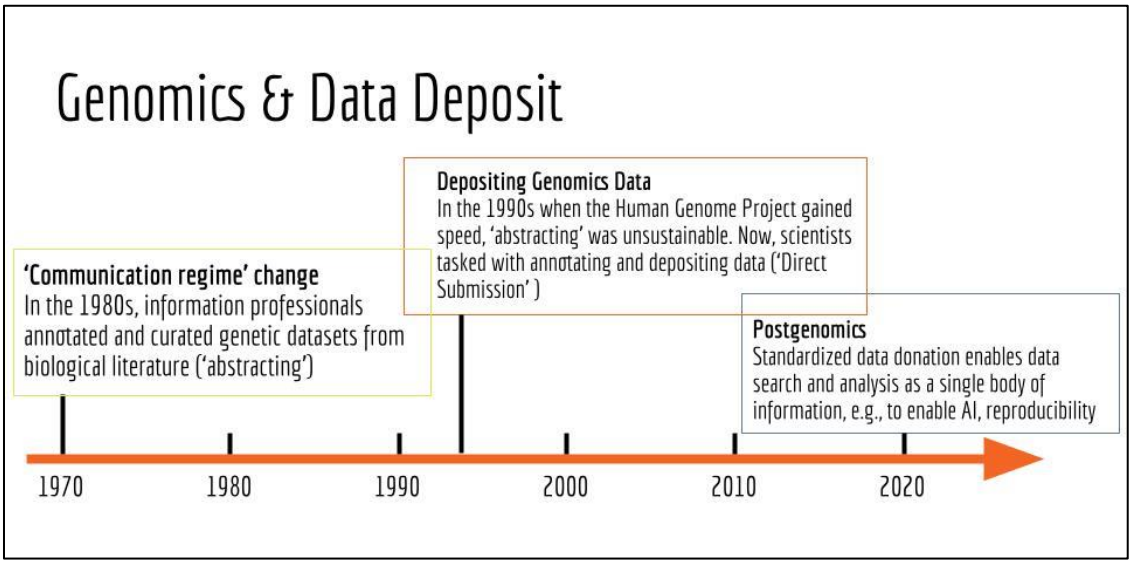

Figure 7: Genomics and data deposit timeline.

The repositories curate nucleic acid datasets and provide products and services that build and grow the submitter/user base (Benson et al., 2017). According to Benson et al. (2017), the data constituting the National Center for Bioinformatics (NCBI) database GenBank primarily includes nucleotide sequences and protein translations. Given the wide use of GenBank and its sister databases in the International Nucleotide Sequence Database Collaboration (INSDC) consortium, the research data included in this study are those represented in these systems. Some of these data are relatively 'simple,' in the sense that they are constituted of four letters

representing base pairs (ATGC), but other data are complex and still are fraught with issues for describing them to capture relevant contextual factors. The purpose of allowing for a wider scope of data is to allow for the emergence of factors associated with data deposit, institutional and those associated with the nature of the data and other factors.

3.5.3 Genomics Data Governance

Early techniques for sequencing the four canonical bases relied on “low-throughput” methods (Durmaz et al., 2015). The last two decades have seen an acceleration of technological and methodological improvement for gene sequencing (Kukurba & Montgomery, 2015). Specifically, from the turn of the 21st century to present, rapid methodological advancement enabled faster, high-throughput quantification of the structure and expression of genes at reduced cost. Sequencing techniques enabled innovation in both basic and applied genetics. Longstanding biological questions could be addressed with the availability of whole genome through transcriptomics, such as the relationship between genetic mutations and environmental factors. Applied areas impacted by these novel methods include biotechnology, forensic biology, and medical diagnosis (Behjati & Tarpey, 2013). Further, rapid DNA sequencing enables a more comprehensive catalog of fully-sequenced organism genomes, leading to proliferation of model organisms and online databases dedicated to them (S. G. Oliver et al., 2016).

Innovations in sequence processing and analysis impacted not only the pace of genetics research but also the epistemology of research design and interpretation of results. Philosopher of science Sabina Leonelli (2014) and critics informed by the philosophy of mathematics argue that big data approaches can easily fall prey to logical fallacies. For instance, Calude & Longo (2017) argue there more data does not necessitate more information: “databases have to contain arbitrary correlations. These correlations appear only due to the size, not the nature, of data.”

Critics argue that big data analysis is vulnerable to the observational bias fallacy (Elragal & Klischewski, 2017; Kitchin, 2014). The ‘streetlight effect’ or ‘drunkard’s fallacy’ describes the logical inconsistency of searching for lost items (e.g., scientific discoveries) where it is easy to find them or where one expects to find them (e.g., sources of data abundance) (Rivera, 2020). The interpretability issues notorious in some machine learning approaches exacerbate the reproducibility crisis (e.g. neural networks) (Leonelli, 2014b). Traditional statistical approaches enabled process transparency because the underlying principles and statistical laws well-known within a community of practice.

As a result, the data-intensive genomics scientific research field has cohered as a community of practice to develop shared professional ethics, terminology, and guidelines for genomics data. Numerous journal articles, white papers, conference proceedings, and other forms of published media exist that provide principles for handling large genomics data. Textbooks have been written on the standard procedures for storing, processing, and analyzing data. There are papers detailed the guidelines for FAIR data management and a “practical guide” for managing large-scale genomic data in research contexts (Tanjo et al., 2021). Even the FDA has created a handbook for industry use of genomic data, with shared vocabulary (e.g., *data, causation, noise)* and principles for good conduct (e.g., “A scientist shall not knowingly engage in cherry-picking”) (*ibid*).

Unlike broader efforts to professionalize data science, the soundness of the approach and epistemological concerns tend to be determined within the disciplinary community of practice (Borgman, 2015; Kitchin, 2014). For example, transcriptomics scientists Kukurba & Montgomery (2015) wrote a review paper of RNA sequencing approaches which cautions readers that a limitation of sequencing techniques is that they require *a priori* knowledge about

the sequences being examined because spurious patterns can appear when highly similar sequences are analyzed; as such, there are threats not only to multi collinearity, but also issues with quantifying genes with subtly different levels of expression (Casneuf et al., 2007; Shendure, 2008). Without disciplinary knowledge and awareness of the vulnerabilities of data and analysis techniques for false positives or negatives, new approaches can raise data validity concerns.

3.5.4 Summary: Institutional Infrastructure for RDM in Genomics

The momentum of large-scale biology and novel sequencing techniques were marked and propelled further by large-scale collaborative projects such as the Human Genome Project (declared complete in 2003) (Hood & Rowen, 2013). Whole genome sequencing also accelerated the creation of completed sequences of multiple species including microbes, plants, and animals (F. S. Collins et al., 2003). Model organisms have played a central part in RDM development in genetics and genomics. Databases dedicated to the collection, curation, and sharing of model organism data have enabled researchers to connect to information and other scientists who are in the shared community. Centralized, web-based databases such as *saccharomyces genome database*, *Mouse Genome Databank*, and *FlyBase* have led to "a bevy of pivotal discoveries that lie at the true heart of the NIH mission" (Hayden, 2016). Yet in recent years, funding cuts have been a cause for an Open Letter to the main funding source, the National Human Genome Research Institute (NHGRI), which supported 5 model organism repositories. The cuts impact the efficacy of niche databases because scientists increasingly volunteer and pay subscription fees to sustain them (S. G. Oliver et al., 2016). Unlike the *GenBank* repository, these niche model organism repositories have less funding and largely lack full-time personnel to consistently perform RDM functions like cataloguing datasets and updating documentation

(Nadim, 2016). Data processing and analysis are one step in the genetics data lifecycle that has advanced substantially in the last several decades.

As a well-resourced, collaborative, ‘big science’ field, the mature knowledge infrastructures in genetics have pushed for greater reproducibility of scientific results. Data governance rules and norms developed out the confluence of multiple factors, both institutional and from the scientific community. The need to standardize data sharing federal agencies who support research data management among other centralized entities, instituted data governance protocols for research data, enforced by publishing companies and reinforced by professional organizations under whose purview data governance efforts were made established common standards (Borgman, 2007; F. S. Collins et al., 2003).

Genetics and genomics knowledge infrastructures have historically benefitted from federal funding. While not without drawbacks (e.g., hyper-competition for grants resulting in risk-aversion (Stephan, 2012; Thursby et al., 2018), funding enables the discipline to acquire the necessary personnel, technology, and other resources important to data curation and the support of long-term data management and community support. Not only do genetics and genomics receive research funding for basic science; the biomedical research enterprise benefits from funds dedicated to science of science research (e.g., SciSIP) (Teich, 2018). SciSIP and similar science of science funding programs can inform process improvement by studying the social and behavioral factors that contribute to productive biomedical and RDM workforces. For instance, U.S. federal funding agencies recently created a funding initiative to support the study of the biomedical enterprise called SciSIPBIO (Teich, 2018).

While genetics is considered a field with mature knowledge infrastructures, prominently those for research data exchange, data sharing in an organized manner is a relatively recent

development as compared to other disciplines such as astronomy or meteorology (Arias et al., 2015; “Data Sharing and the Future of Science,” 2018). Now, in genetics, genomics and structural biology, shared datasets are common (e.g., EMBL). Researchers have used and re-used previously published datasets to enable new discovery in these areas (Leray et al., 2019). The core facility scientists have played an increasing role in data production, rather than faculty themselves. The culture of data sharing through deposit to online repositories has grown with institutional measures to protect intellectual property and facilitate data documentation. Moreover, sequence data is simply structured and has become more standardized (Now, 2016).

However, representing the context of biosamples and metadata for diverse collection sites remains a gap, e.g., metagenomics (Bietz & Lee, 2009). This is similar in astronomy, where “data are inseparable from the software code used to clean, reduce, and analyze them” (Borgman, 2015, p. 106). Overall, genetics and genomics knowledge infrastructures are composed of “robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds” (Edwards, 2010). The continuity of disciplinary expertise is enabled by a strong information ecology for RDM, with systems to implement information exchange, preservation, searchability, discovery, and sharing.

3.6 Social Science Research Data Management

Social science research is the study of human behavior and social processes. For example, social scientists study individuals and populations to understand their experiences. The research environment in the social sciences is less institutionalized in terms of data infrastructure for widespread sharing, especially in comparison with the genomics context. However, it is not as if the social sciences have no institutional infrastructure for data management and deposit. Institutionalization measures, cultures of sharing, and have policy mandates do exist – in fact, the

Interuniversity Consortium for Political and Social Research (ICPSR) was initiated in 1962, almost two decades before the flagship genomics repository (i.e., GenBank) was started at the Los Alamos National Laboratory (1979). (Benson et al., 2017).

As a research site, ICPSR is an excellent space to examine a less-institutionalized context (compared with genomics research data management), but one in which there are growing RDM guidelines, policies, standards, and norms for research data. In the social sciences, there exist infrastructures to manage and deposit data (e.g., repositories such as ICSPR). However, there is still not a lot of pressure to deposit data, in part because of the challenges with de-identification of data to preserve confidentiality. As a result, research data management can be based on in-house, ad hoc solutions for creating RDM structures and shared norms. Here, there are also context-specific incentives e.g., funding (i.e., NIH) and publisher requirements. As well, there are some metadata standards, e.g., specific to a repository. However, there are sometimes a lack of shared data quality principles, e.g., specific to professional associations.

3.6.1 Social Science Data

Social science data are observations of social phenomena, represented by a variety of data types, formats, and collection and analysis methods (e.g., discourse analysis of interview transcripts). They include confidential information, and often human subject data is represented at an individual (i.e., non-aggregated) level. As such, social science research data are sensitive; protecting them is an ongoing challenge in data sharing efforts in the social sciences. They include topics such as youth behavior (Ahlin, 2020), prison recidivism (Wang et al., 2010), and social epidemiology (Campos-Mercade et al., 2021). Research data in the social sciences are often represented as text, as in qualitative interview or observational data, but also can include

images, video, or other media or objects. As well, social science data are quantitative if collected, e.g., via survey or experiment.

3.6.2 Social Science Data Repositories

The role of open data repositories in the social sciences is like that of genomics or other fields: to gather and standardize qualitative and quantitative data into a federated source, enable the contents to be searched and analyzed as a single body of information, and for researchers to share datasets for further analysis more widely. They also include services and products to support researchers with their data (e.g., data analysis tools, educational materials, workshops). Databases also include policy to protect researcher's data from being scooped (e.g., embargo policy). Documenting methods to promote transparency are also a function of databases in the social sciences. Examples of open online data repositories include ICPSR, Dataverse, and Figshare. These are distinct from open online government data repositories, such as the U.S. census, or other social and economic data collected by federal agencies for population statistics, policymaking, and defense intelligence.

The timeline of the social science data repository ICPSR is an exemplary model of the technical features and historical development of social science repositories (**Figure 8**). ICPSR is a data repository that started in 1962 and has become well-established as an authoritative place to deposit and access research data. It gained prominence and partnerships with federal agencies such as the National Institutes of Justice (NIJ), which has policies requiring that funding recipients deposit their data in ICPSR. ICPSR's development is telling as to the extent of institutionalization and the ways it gained legitimacy as a 'go-to' repository for social science data. The original repository was a political research repository in the 1960s, changing its name to include the social sciences in 1975.

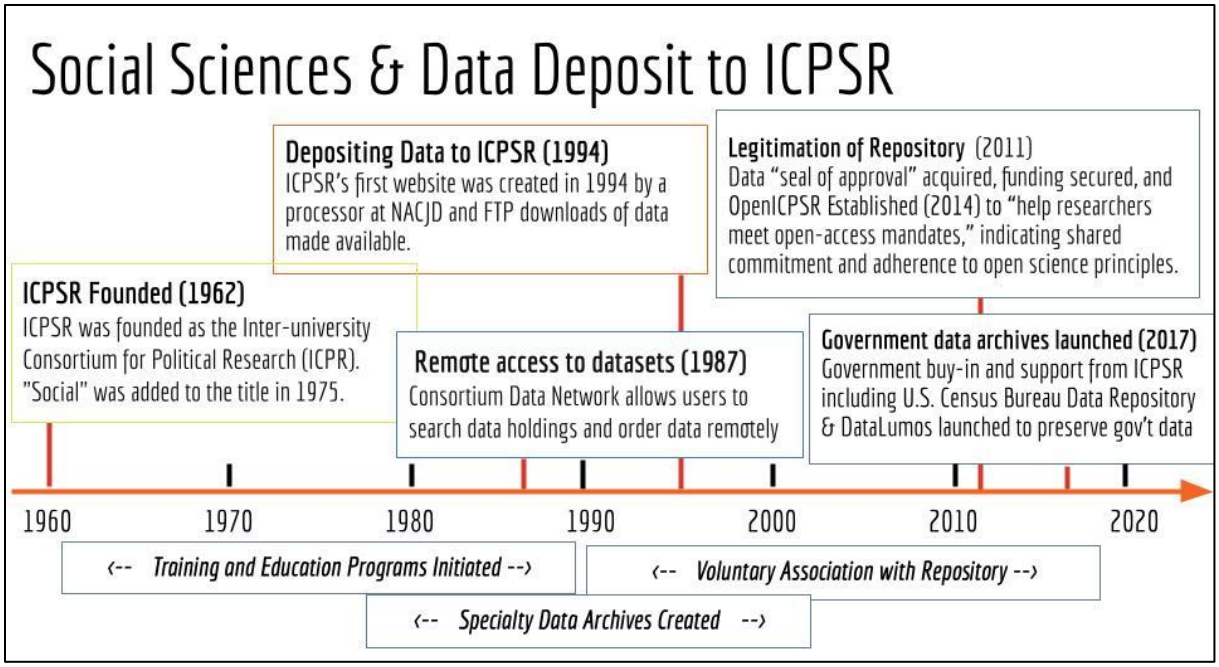

Figure 8: Social science research data repositories: ICSPR timeline.

Data deposit in the social sciences can often be a labor-intensive process. It can also be expensive for the submitters, an activity not directly compensated for (explicitly) through grant funds. Some of the key activities to deposit data to ICPSR, for example, include locating the data and naming the files, then proceeding to de-identify individuals (e.g., redacting participant names, organization titles, or other potential information that could breach confidentiality in an interview transcript). The datasets must then be annotated to describe what variables are contained therein, the methodology used, and the population demographics. In addition to this de-identification and annotation, researchers are also required to submit documentation for the datasets that further detail the data cleaning, processing, and organizing steps. Once the researcher completes all the required steps, she uploads the data to the repository staff for their review and approval, a process which can take months.

Once approved, the data is then available on the ICSPR website and searchable. Some of these datasets are classified as "open" data; that is, they are immediately available for download

to anyone who can access the webpage for the dataset. Other data requires signing in or registration through the ICSPR portal. Overall, the data deposit process in ICSPR and other social science repositories (e.g., Harvard Dataverse) necessitate work to ensure the data are well-described and do not breach participant privacy. The infrastructure for deposit has developed substantially over the past 60 years.

3.6.3 Social Science Data Governance

Governance of social sciences research data is shaped by the policies and platform rules for data. There are a few widespread policies about how research data should be stored and shared, largely determined by the Institutional Review Board (IRB) of a university or other institution. For example, IRB requires specific care taken with human subject data to ensure confidentiality and protect participants. The policies for data sharing are often also specified by individual institutions, e.g., the academic institution. For example, Memoranda of Understandings (MOUs) signed by the parties sharing data (e.g., researchers at university A, college B, and company C) elaborate and dictate how data will be shared, with whom, and under what circumstances. The data management specifics are not widely shared beyond these core issues, i.e., that of data storage, security, and inter-organizational sharing. The faculty can manage them as they choose and use the software tools and analysis method(s) they prefer within these governance structures.

Governance of data deposit varies widely. The main pressure social sciences researchers face in depositing data are from funding agencies and journal publishing. Some of the prominent funders in the political sciences, such as those in criminology, do require data deposit from the federal agencies. For example, a prominent federal agency requiring data deposit is the National Institutes of Justice (NIJ). They specifically require grantees to deposit with ICSPR. Several

other federal funding agencies also partner with ICSPR, making them their de facto repository for grant recipients to deposit to. Less common, but still a source of pressure to deposit data, are foundations that fund social science research. They tend to be less specific about where the data should be deposited, but nonetheless sometimes include deposit as part of the funding contingency.

Much less formally, there are several best practices for data management and deposit. In the social sciences, these tend to be tied closely to methods and the domain, and therefore somewhat discipline specific. Best practices for data come a variety of sources, including qualitative methods textbooks, professional organizations, and the apprenticeship-like experience graduate students have in their doctoral research training. For example, professional organizations hold conferences that include workshops on methods, and within this topic, can include incidental references to examples, suggestions, and recommendations for how to manage data. These best practices, within each specific social science discipline, compared to genomics, tend not to be codified, e.g., as a professional standard.

A notable exception to the lack of codification is the data transparency statements and requirements of the American Association for Public Opinion Research (AAPOR), who has been increasing its efforts to make sure researchers document their data and methods rigorously, due to skepticism about the validity of political polling from the public and scholarly community. The AAPOR is an example of an organization that has led to a social science sub-discipline establishing shared best practices and governance for RDM.

3.6.4 Summary: Institutional Infrastructure for Social Sciences RDM

The institutional infrastructure for data management and deposit in the social sciences has been growing over the past 60 years. The ICSPR data repository has had the trappings of

institutionalization since its inception. However, as it grew, received funding and built its leadership, it gained legitimacy. For example, it has key characteristics of becoming institutionalized, including offering training and education that was taken by researchers who deposit data, a website (1994), and a data "seal of approval" from the governing body. It also showed signs of legitimacy when government data archives were launched in 2017 (**Figure 8**).

However, it has not accelerated at the same pace as other fields, e.g., genomics. The pace of genomics data exchange was accelerated by various factors, including the Human Genome Project, the conductivity of the data for sharing (e.g., the comparative ease of DNA sequence data – nucleic acid base pairs comprised of 4 letters), and applications in resource-rich sectors such as medicine and biopharmaceuticals.

As such, the social sciences do not have as widespread tools compared to NCBI, which has developed a suite of analytic tools, data visualization and manipulations, and training workshops on how to submit data. Representative repositories such as NCBI's GenBank and ICPSR both have a full-time staff, but on very different scales. In addition, pressure is not as strong for institutional pressures (e.g., from Journals and funding agencies). Overall, the institutional support and pressure is not as pervasive in the social sciences as it is in genomics. Support for data deposit, as well, is not as pervasive.

The research setting of U.S. academic institutions, with a focus on research-active faculty who deposit data, sets the context for the research design as well as the qualitative methodology used. The setting allows for a comparison between genomics and social sciences research data management, described further as part of the methodology in what follows. Specifically, in the next section, I describe the research methodology of the dissertation. Included are the methods (for data collection, analysis, and confidentiality and data security) used across all three studies.

3.7 Research Methodology

In this section, I describe the methods used in all 3 studies/papers: criteria and sampling method for selecting the population (i.e., purposive sampling of research-active U.S. faculty who have deposited to a data repository), recruitment methods (i.e., email, phone call, and snowball sampling), data collection (i.e., semi-structured interviews that were audio recorded and later transcribed), data storage (including confidentiality and data security), and data analysis (i.e., content analysis, inductive and deductive, drawing from grounded theory and the constant comparative approach). I explain and justify the method selection choices and note the methods I considered (but ultimately did not select).

3.7.1 Sample Selection

Non-probabilistic purposeful sampling technique was employed to select the samples across all three studies. This technique is used to recruit participants with predefined characteristics that are relevant and informative for addressing the research questions (Tashakkori et al., 1998). The participants selected for this study were U.S. faculty in academic institutions and had no overlap for the three studies (Study 1, Study 2, and Study 3). In other words, data was collected from three separate samples of faculty. All participants are research-active, tenure-track academic faculty at R1 (of the "Carnegie Classifications," administered by the Carnegie Foundation for the Advancement of Teaching and the American Council on Education (ACE) academic institutions in the United States. Study 1 participants (n = 12) were constituted of molecular biology faculty. Study 2 participants (n = 18) were constituted of faculty in genetics and genomics. Study 3 participants (n = 15) were constituted of faculty in the social sciences who deposited to ICPSR.

Table 6: Participants in each study and sample size

	Study 1	**Study 2**	**Study 3**
Population	Faculty: molecular biologists in genetics and genomics	Faculty: genomics	Faculty: social scientists
Sample size	n = 12	n = 18	n = 15
Purpose	Explore the experiences of data deposit	Develop a theoretical framework of data deposit	Apply the theoretical framework to broader context

The sample selection was purposive, to sequentially build on the previous explorations of the data deposit experiences, attitudes, and practices of scientists to explore, analyze, and develop a theoretical framework of data deposit in a rising environment of increasing institutionalization of data deposit. Initially, Study 1 also included research scientists in academic institutions and graduate students. The division of labor on data work proper includes core facility researchers, and research data centers at the academic institution, as well as graduate students, postdoctoral researchers, and research staff. The inclusion of non-faculty was motivated by prior literature suggesting the data work in genomics is not done primarily by the faculty PI (e.g., Ankeny & Leonelli, 2015) and that it would be an oversight to not include the "adjuncts" of genomics data work in the initial phase of the exploratory study, given its goal to explore and uncover a range of issues related to data work (Scroggins & Pasquetto, 2020).

However, as study 1 progressed, the population narrowed to only faculty in genomics. It became clear that to answer the research questions about how research data work is organized to deposit data, and the extent of institutionalization of data management at the lab level, faculty are the center of decision-making. My focus turned to the supervision and organization of data management in the lab/research group. Faculty have central purview over this unit. As such, they

are the ‘front lines’ of how policy is translated into practice because they control of how the lab manages data. For example, they set the plan for how data management workflows unfold. They decide the tools and software to use to deal with data storage, processing, and analysis, key steps in the data management process. Importantly, faculty also set expectations about how their lab members should deal with data (e.g., collect, store, analyze, and share data).

For these reasons, faculty were the selected population for assessing the maturity of the data management processes in a research group – faculty they have the ‘big picture’ view of their project and its data. They can speak to their experiences of the institutional pressures to manage data – the nature of the faculty job is to maintain close contact with professional communications, events, and news that refer to data standards, mandates, and disciplinary norms about the existence of and expectations for data management. Faculty members supervise the lab, including students at the bench, and are responsible for training students, implicitly or explicitly conveying how data should be managed. Even if training and supervision is delegated (e.g., to a postdoc or a lab manager), the faculty PI nonetheless writes the key aspects of handling data, since they have a stake in data quality, accessibility, and the like. This justification led to a culling of only faculty as the populations for Study 2 and Study 3.

The criteria for selecting the faculty in this sample were research-active faculty at R1 U.S. academic institutions. The faculty sample comprising the data in studies 1, 2, and 3 are from R1[4] institutions (of the “Carnegie Classifications”). The Carnegie ranking implies some characteristics about the institution relevant to the level of institutionalization of data management, e.g., level of support for data (e.g., the presence of research data centers (RDCs).

[4] As of the writing of this dissertation, “R1” is a Carnegie Classification category indicating doctoral degree granting, “very high research activity” institutions.

The disciplines of the faculty sample are, broadly, genomics and the social sciences but include subdisciplines given the faculty research interests and training (e.g., biophysics, social epidemiology). Study 1 focused on molecular biology faculty who do empirical genetics and genomics research. Study 2 also focused on genomics faculty, but was also inclusive of biochemistry, a closely related discipline that also employs genomics approaches. Study 3 selected for social and political science faculty who submit data to ICPSR. A focus was on qualitative approaches, but quantitative methods were included to allow for the possibility that things like data type and level of aggregation were factors associated with the institutional maturity and pressures impacting on data management and deposit practices.

3.7.2 Recruitment

Recruitment proceeded by contacting participants who fit the purposive sampling criteria. Recruitment scripts were developed and included an email and phone invitation script, a follow-up email script and a thank you for participating email script. The scripts saved time during recruitment and served to maintain consistency across participants. Participants were then contacted via email or phone. If participants did not respond the first time, one follow-up email and/or phone call were done. Once participants demonstrated interest, the consent form was sent for their review and a time for the interview was scheduled.

3.7.3 Data Collection

The data collection proceeded through interviews. Interviews are a rich qualitative collection instrument, probing for unexpected topics and deeper understanding about the topic at hand. The semi-structured “conversational” interview is effective to gather data on experiences, opinions, perceptions, and feelings (Creswell & Poth, 2016). Additionally, they afford contextual understanding of the preferences, attitudes, and knowledge that inform behavior. The process of

interviewing as a method involves the careful design of questions, the definition of a population, the expected responses, and the reason for pursing interviews to acquire in-depth information.

An ethnographic approach was considered for the study, given its use in traditional social studies of science. Ethnography is a research method emerging from anthropology that is useful for the study of human culture, including work and social dynamics. In the social studies of science such as Science and Technology Studies (STS), scholars employed ethnographic examinations of scientific work in laboratories. Canonical studies representing this body of work include “laboratory studies” (Ziewitz & Lynch, 2018) such as Latour & Woolgar (1986), among others. An ethnographic approach was considered for this study, given its utility in the study of scientific culture and practice. The main benefit an ethnographic approach would have afforded this dissertation would have been to observe the practices of researchers. Studies of practice often benefit from direct observation, given the biases associated with self-report (e.g., individuals are notoriously bad at recounting or remembering the past, social desirability bias, acquiescence bias) (Bobko et al., 2014; Creswell & Poth, 2016). However, interviews were selected for practical and theoretical reasons including to elicit attitudes and perceptions, the limitations of in-person data collection during the COVID-19 pandemic, and account for the need to limit time-intensiveness of the study.

First, the data required to address the overarching research questions included the attitudes and perceptions of researchers. Interviews are an excellent approach for eliciting data related to beliefs, attitudes, and perceptions. Second, also for practical reasons, the COVID-19 pandemic complicated use of an ethnographic approaches. Observing individuals in-person would have been not only more challenging because of scientists began to work from home, but also would be ethically questionable to potentially expose the participants to the risk of the virus

(i.e., by interacting with the researcher in a traditional laboratory setting). Third, interviews were less time-intensive than ethnographic studies. Ethnographic are time-intensive because they require not only the time spent immersed in the lab but also to build trust and relationships with the study participants.

The choice of interviews over ethnography impacted several aspects of the study. Importantly, selecting interview as the method scoped and shaped the research design and impacted what data could be collected and what data was necessarily excluded. As mentioned above, the selection of interviews emphasized the attitudes and reported practices of researchers but precluded the direct observation of behavior to triangulate this data. Further, the choice to start with interviews as the method of data collection (and content analysis) in Study 1 trickled down to the rest of the studies. This resulted in a few limitations in the rest of the studies because the findings that resulted from the interviews were used to develop the framework, a framework that was subsequently tested in the final study (i.e., Study 3).

A limitation of study 1 is that it scoped the remaining studies to the themes surfaced initially. The work was limited by this initial "seeding", that is, beginning by defining the scope of the phenomenon of Interest to data practices that led to deposit; plus limiting the remaining studies to the themes surfaced in Study 1. However, study 2 and study 3 are not strictly limited to the themes which study 1 surfaced. I allowed for additional categories, themes, and codes to emerge during coding. The limitations of beginning with Study 1 codes and themes to develop a theoretical framework, then, did not close off the possibility for new themes to emerge. The sequential nature of the data collection did sensitize the research to the initial themes surfaced, which limits the generalizability of the study findings. In conducting the interviews, I also visited

labs, made observations and fieldnotes of the laboratory workspaces, lab animals and equipment, meeting and office spaces, and signage that appeared to be part of the data workflows.

While interviews have elements of ethnographic research, given my immersion in the laboratories, occasional lab tours and relationships I built with some of the participants, it is not a formal, systematic ethnographic study. Ethnographies involve preparation by the researcher to immerse and become a participant-observer, being involved in the research site long enough to become 'native' to the environment. In fact, ethnographic or phenomenological research, this method can be especially valuable for eliciting the *meanings* of themes which occur in the "life world" of participants, given their subjective point of view. Interviews were ultimately selected, then, because they are appropriate for design of each of the three studies (i.e., phenomenological, grounded theory, case study, respectively). Ethnographic studies also draw from multiple sites. In addition, the time constraints and the COVID-19 pandemic limited my ability to immerse for extended periods of time in the scientists' labs and with their lab groups.

Interview audio was recorded and transcribed. I used a non-cloud-based handheld audio recording device when conducting interviews in person. If the interview was conducted via video conference (e.g., Skype, Zoom), the audio was recoded using the video conference audio recording feature. Transcription software was used for initial speech-to-text transcription, specifically: Rev.com, Temi, and Zoom's transcription software.

3.7.4 Data Analysis

Across the three studies, I employed content analysis and the constant comparative method for qualitative data analysis (described in each chapter in greater detail). *Content analysis* is an analytic method that is conducted in the study of document, content analysis is a useful approach for analyzing communication artifacts from transcripts to historical documents. The philosophical and practical approaches to content analysis vary between disciplines, but all

involve the “systematic reading or observation of texts of artifacts which are assigned labels [codes]…to indicate the presence of…meaningful pieces of content” (Saldaña, 2015). Statistical methods or qualitative methods can then be employed to analyze the content. For instance, thematic analysis “emphasizes pinpointing, examining, and recording patterns (or "themes") within data” (Braun & Clarke, 2006). The patterns across texts are called themes. I used a qualitative approach to identifying themes across the texts.

The *constant comparative* method is used by researchers to develop concepts from the data by coding and analyzing at the same time (Kolb, 2012). To do this, the researcher “combines systematic data collection, coding, and analysis with theoretical sampling in order to generate theory that is integrated, close to the data, and expressed in a form clear enough for further testing” (Scott et al., 1993, p. 280 in Kolb, 2012). According to Glaser & Strauss (2017), the constant comparative methodology has four stages: “(1) comparing incidents applicable to each category, (2) integrating categories and their properties, (3) delimiting the theory, and (4) writing the theory” (p. 150).

Study 1 and 2 employ *inductive* content analysis, drawing from grounded theory, to generate a theory explaining data deposit (i.e., the “data articulation” framework). Inductive content analysis is used when there is prior empirical work addressing the topic or when it is fragmented (Elo & Kyngäs, 2008a). Study 3 used *deductive* content analysis, applying the theory (generated in Study 2). Deductive content analysis is useful in cases where the general purpose is to test a previously generated theory in a different situation. A deductive approach uses operational measures based on prior knowledge to structure the analysis (Elo & Kyngäs, 2008a). A subset of these criteria is summarized in **Table 7**.

Table 7: Summary of the sample population discipline, recruitment techniques, and data collection/analysis approaches compared in each of the three studies.

Study	**Sample (discipline)**	**Recruitment**	**Data Collection** (Interviews)	**Data Analysis**
Study 1	Genomics, genetics	Email	In-person, phone	Inductive content analysis
Study 2	Genomics, biochemistry	Email, phone	In person, phone, zoom	Inductive content analysis, drawing from grounded theory
Study 3	Genomics, social and political science	Email	Zoom	Deductive, constant comparative method to allow for emergent themes

The data analyzed were interviews transcripts and documents (Study 3 only). Study 1 and 2 analyzed interviews only. In study 3 I performed content analysis on documents collected from participants, e.g., lab handbooks, data archiving plans (described in greater detail in study 3 methods section). A coding scheme was developed based on the themes that emerged in analysis. Accordingly, Study 1 and Study 2 employed inductive approaches to developing the coding scheme and an inductive-deductive approach in Study 3. I drew from grounded theory most heavily in Study 2, as it was the study in which the theoretical framework was generated.

The issue of using multiple coders in qualitative data analysis (QDA) is an unsettled topic of debate. Some argue it is unnecessary, and that is a redundant practice that comes from trying to mimic quantitative studies (O'Connor & Joffe, 2020). Other argue it is necessary to avoid bias and increase "data quality" (Church et al., 2019). Using multiple coders in QDA has advantages and disadvantages, reducing bias and maximizing conceptual clarity in codes where there is disagreement (e.g., due to conceptual muddiness) ("Spotlight on Qualitative Methods," 2020). A disadvantage, however, is that the push to agree across all coders can come at the expense of "interpretive insight," where the unique perspective of a coder can add insight to the analysis that would otherwise be lost in agreement. In this study, Study 1 did not use multiple coders, but I employed methods associated with QDA rigor including researcher reflexivity, constant

comparative approach, reflective memos (Creswell & Poth, 2016). Study 2 included a second coder. We discussed the codes to converge on an agreement about their meanings and identify ill-defined codes to produce greater clarity.

The reason for using content analysis, rather than another analysis approach such as discourse analysis is because it is a non-invasive method of pattern identification and examination which can be replicated in systematic ways to uncover the object of interest in the study. Content analysis involves sampling from a source and commonly used if intending to generalize results to a broader population (Bauer, 2000), which was appropriate for the general aim of this study, to generate a framework to explain the work behind research data deposit to open online repositories.

3.8 Confidentiality and Data Security

Confidentiality and data security were assured across all three studies. The actions taken to ensure confidentiality of participants identity and information were concerning security, storage, and responsiveness to participant requests. Specifically, where zoom was recorded, the videos were deleted at the participants request. Otherwise, the videos of the call are preserved to allow for the data included in screen sharing to be available for analysis (e.g., I shared my screen to show the ICSPR dataset the faculty submitted to allow the faculty to discuss the deposit process). The updated contact information was entered into a spreadsheet and stored securely. The transcripts were de-identified and anonymized and stored on a secure server. The device used for recording is kept in a locked cabinet in a limited-access lab on campus.

3.9 Chapter Summary

In this chapter, the overall research design was described. The shared aspects of the methods were explained, including the research setting and methodology. The research setting

section described individually and then compared the research environments of genomics and social sciences researchers (the sample populations) with respect to their current and historical institutional pressures and practices related to research data management and deposit. I summarized the shifts in both fields and the academic culture and technological issues that shape the data management and deposit environment (e.g., data deposit is not 'glorified' in traditional academic crediting culture).

This description of the research environments for RDM re-emphasizes the central research questions of this dissertation study: Why do faculty still submit data (if it is not glorified, e.g.)? While the literature shows it is because of institutional pressures (e.g., journals, funding agencies, and professional norms, disciplinary mandates), it is not clear, given these institutional pressures, what "makes it possible" to deposit data – there is also a need for workforces – faculty, staff, core facility researchers – they do the local work to align institutional goals, and to create structures where there are none. Indicators of institutional support is if there is funding supporting hiring and lab activities, that is concrete way that policy reinforces the necessary work of data management. The next three chapters examine these questions systematically.

CHAPTER 4
STUDY 1

This chapter reports an exploratory study that sought to investigate data practices of biosciences faculty in depositing data to a data repository. The study employs a qualitative semi-structured interview approach with U.S. academic research faculty in molecular biology and genetics at R1: Doctoral Universities – Very high research activity (in the Carnegie Classification of Institutions of Higher Education) (n = 12). The purpose of the study is to address the gap in domain specific studies of the early stages of the research data life cycle. The study contributes to the scholarly communication literature by situating large-scale bibliometric studies of biomedical data practices (e.g., Cronin & Sugimoto, 2014; Larivière et al., 2016; Teich, 2018) in context among the professional practices, organizational cultures, and disciplinary norms of data-intensive academic research in the biosciences.

The main contribution of the study is the identification of eight analytic categories (i.e., theoretical constructs) related to RDM and data deposit practices within data-intensive practice: Administrating and Managing, Maintenance and Repair, Collaboration and Relational Labor, Archiving and Documentation, Socialization into Data Culture, Data Articulations, Publishing Activities, and Risk and Uncertainty Management. The analytic categories are articulated in relation to the literature on corresponding theoretical constructs, providing the basis for developing the research framework explaining data deposit.

This chapter is organized into three sections. The first section describes the rationale and goals of the exploratory interview (Rationale & Goals). Section two reports the study design, guiding research questions, and findings in association with the relevant literature (Exploratory Study). The chapter concludes with a discussion of the contributions and limitations of the

exploratory study and a chapter summary (Conclusion). The overarching purpose of this chapter is to lay the empirical and conceptual groundwork for creating a research framework to explain the labor of data management to be tested in a following study.

4.1 Rationale & Goals of the Exploratory Study

Substantial investments in cyberinfrastructure (CI)-enabled science have led to a flurry of interest in effectively supporting data-intensive biosciences. The topic is of concern to a range of stakeholders, from science policymakers and academic research faculty to academic library professionals. Yet despite the substantial federal investments, a paradox characterizes the conceptual and empirical landscape of data practices: 'abundant practice, absent theory.' That is, there is widespread practice of data management and deposit but a lack of conceptual frameworks to analytically describe the types of practices and scholarly content of research data management (RDM) and deposit work.

Given the relatively nascent literature in faculty research data management leading to data deposit and this paradox of abundant practice and little theory, this exploratory study was conducted to identify and describe the data activities of U.S R1 academic faculty within the research cycle. The overarching research question of the study is: What are the experiences data deposit for faculty in genomics? This question is further broken down. Specifically, to identify the data practices faculty within data management research life cycle, the first research question of the exploratory study asks: *What are biosciences faculty data management practices*? (RQ1)

Overlooking data management practices under the purview of faculty implicitly neglects a component of the RDM quality and value chain. Neglecting a component of the value chain can undermine the entire chain (Porter, 1985). From an organizational standpoint, ignoring how faculty make data management decisions and their perceptions and priorities can misrepresent

their work, and create friction between researcher desires and institutional decisions (Barley & Tolbert, 1997; Feldman & Pentland, 2003). Simultaneously, is unclear whether faculty RDM can benefit as a site for faculty development, process innovation, or as a place potentially generative of best practices. To address this empirical gap, the second research question of this study asks: *How do faculty make data management decisions*? (RQ2)

4.2 Methods

Exploratory studies are useful for developing and refining theory (Berg and Lune 2012) and can be especially useful for clarifying and refining theoretical constructs and definitions (Pickard, 2013). As such, this exploratory study examines the experiences, perceptions, and data practices of bioscience faculty. The exploratory study has a formative function to the overall dissertation study. In this section, the target population and recruitment are described and the methods for data elicitation and analysis are overviewed.).

4.2.1 Study Scope

Semi-structured interviews were conducted with R1 academic faculty (n = 12, 5 Female, 4 pre-tenure) to explore faculty's experiences and perceptions of data practices, specifically, the disciplinary culture, the main actors and actants, and the artifacts, norms, and practices involved in data-intensive work. The interview questions were designed to elicit scientists' experiences around data management and deposit and surface the data activities, perceptions about RDM, and data cultures of the research group and discipline. Although the interview questions and analysis focused on data management and deposit practices, the unexpected themes and topics surfaced. Because of the exploratory nature of the study, these emergent categories did not pose a significant challenge, but did impact on the scope of the results exceeding the scope of the initial study design.

The next sections describe the target population, the participant recruitment process, data collection, data analysis methods, and findings. I conclude with a brief discussion of the emergent themes and categories to establish the main factors to be developed in future research.

4.2.2 Target Population

The target population was selected through establishing *a priori* criteria and verifying these criteria with manual web searches and by extracting metadata records from scientific data repositories in data-intensive biology research-active academic faculty in Molecular Biology. All were recruited from Carnegie Classification "R1" Universities. A range of academic ranks, gender, and biological sub-disciplines were selected to provide an approximated representation of study population, given the exploratory purpose of the analysis. Initially, preferred participants were those who had submitted to the GenBank repository to use the submitted metadata records as a cultural probe to confirm our interpretation of the meanings of the metadata fields. However, there records contain name ambiguity which made it difficult to correctly identify faculty (Qin et al., 2015). The target population was further narrowed by geographical proximity to control for state policy and to facilitate site accessibility (in the case of a follow-up site visit). The decision to focus on these factors for the targeted population was to acquire information from faculty who have used GenBank or are biological scientists.

Molecular Biologists in academic institutions have community-belonging to various types of research communities, according to their topics (biofilm, fish pathology, reproductive evolution), methods (field experiments, observational studies, simulations, proteomics), and/or material (e.g., model organisms, reagents). Their research frequently spans multiple research disciplines based on the methodological or domain expertise needs, such as when an omics researcher's robotics or sequencing machinery is needed for a neurobiology. In short, there is a

wide range of topics and approaches within Molecular Biology, but all leverage scientific methods across research paradigms from observational and experimental to simulation-based and computational using theoretical and empirical approaches. In general, biologists' expertise is to contribute knowledge of the processes and dynamics of living systems through empirical and theoretical approaches.

The target criteria resulted in the study population. The study population selected was faculty within the research community of Molecular Biology who focus on genomics and data-intensive methods (e.g., proteomics, metagenomics). The purpose of selecting this population was in pursuit of the goal to explore the contexts of scientific data repositories use and explore the utility of academic research laboratories as a research site. The population was broad enough to elicit diverse responses but specific enough to support the research goal: to examine the cyberinfrastructure-enabled data-intensive RDM and data deposit practices among academic faculty. The GenBank repository is one of many databases used by the genomics community. Therefore, individuals not using GenBank but using other types of databases were still included and recruited to the study. Researchers were recruited from both public and private R1 universities which also provided a diversity of teaching, research, and service expectations. In sum, the target population was identified first by the self-identification by scientists whose disciplinary expertise on their professional or department webpage indicates they are in the genetics and genomics community with a secondary criterion using the manual web search approach was how the department classified the faculty.

4.2.3 Recruitment of Participants

The target population was recruited through three strategies to select for prospective interview participants: GenBank metadata record search, web search using academic faculty

rosters, and snowball sampling. This approach is a non-probabilistic purposive sampling technique. The technique is used to recruit participants with predefined characteristics that are relevant and informative for addressing the research questions (Tashakkori et al., 1998).

The first recruitment strategy was to identify potential participants most informative for gaining information about the GenBank community, specifically. Thus, this recruitment technique enrolled participants by identifying in GenBank metadata. The metadata fields of "author name," "year," and "description," were extracted and the data mined for identifying information to recruit faculty researchers. The metadata mining method uses public records available in the NCBI data repository. Though a newer technique within Internet recruitment methods, Hine (2005) identifies bibliographic metadata mining as a type of public documents that is valid, ethical, and reliable as an approach to Internet research recruitment. The second recruitment strategy was web search[5]. This is a common strategy in organizational studies of scientific work (e.g., Thursby et al., 2018) and in studies of scholarly communication (e.g., the use of publication metadata and CVs by Katz, 1994; Simonton, 1997). The faculty identified in GenBank metadata were then verified as "research active faculty" through the web search. Updated contact information was entered into a spreadsheet and stored securely.

The third recruitment strategy was snowball sampling. Drawing from the pool of initial interviewees and through my social network of contacts, additional participants were selected as prospective participants. Snowball sampling (also referred to as link-tracing or chain-referral in social network analyses) is a sampling method driven by the participants (Fehr et al., 2018). A limitation of snowball sampling is it requires knowledge of 'insiders' in a population (prior to

[5] The web search relied on Google's search algorithm (2018) and the university faculty rosters in Biology departments focusing on genetics and genomics.

data collection) (Atkinson & Flint, 2001). Nonetheless, the targeted population is an accessible population. That is, compared to "hard to reach" populations such as drug lords and refugees, this exploratory study identified "insiders" using the respective academic faculty directory and recent publications to select participants for follow-up interviews. This study drew from the snowball sampling procedures of Illenberger & Flötteröd (2012) and (Gile & Handcock, 2010). Participants were contacted via email. The participant demographics and disciplinary-affiliation details can be found in **Table 8.** An ID anonymizes the individual participants, their position indicates their seniority, as well as if they have attained tenure (an assistant professor title indicated pre-tenure and others are tenured). Participants are from departments that lie in the areas of Molecular Biology, Genetics, and Biochemistry.

Table 8: Study 1 Participant Demographics.

ID	Gender	Position/Title	Discipline
P1	Male	Professor	Biochemistry & Molecular Biology
P2	Female	Professor, with a leadership role at the department level.	Biochemistry & Molecular Biology
P3	Male	Professor	Biochemistry & Molecular Biology
P4	Male	Professor, with a leadership role at the university level.	Biochemistry & Biophysical Sciences
P5	Male	Assistant Professor	Biochemistry & Molecular Biology
P6	Female	Associate Professor	Molecular Biology & Genetics
P7	Male	Professor	Biochemistry & Molecular Biology
P8	Female	Professor, with a leadership role at the department level.	Molecular Biology & Genetics
P9	Female	Assistant Professor	Molecular Biology & Genetics
P10	Female	Associate Professor	Biomaterials
P11	Male	Assistant Professor	Biochemistry & Biophysical Sciences
P12	Male	Assistant Professor	Molecular Biology & Genetics

The invitation e-mails were extended to individuals identified through all three recruitment strategies, with 2 from the GenBank metadata mining technique, 9 from web search,

and 1 from snowball sampling. Three rounds of emails resulted affirmative responses to a 40 to 60-minute interviews to discuss data practices and the context of their work, more broadly.

4.2.4 Data Collection and Analysis

From April 2018 - August 2018 semi-structured interviews were conducted using an interview protocol to guide the questions (see Appendix A for interview protocol). The questions were designed uncover factors related to data management and deposit practices, the organizational and data culture of Molecular Biology research, the historical factors associated with practices, such as technological development, ethics of information sharing, and workflows of typical data-intensive projects.

To gain an in-depth description of individuals' experiences, a phenomenological approach was employed (Moustakas, 1994). Phenomenological perspectives assume that meanings are emergent properties of sociotechnical contexts, and the meanings individuals ascribe to the technologies can be uncovered by self-reported behaviors and perceptions. A phenomenological approach was determined as appropriate because of its strength in eliciting the experiences of people, in this case, in depositing data to a repository. The interview question protocol design reflected the phenomenological approach and goals.

The interview questions were structured in a funneling approach, starting with broad questions about how Molecular Biology faculty organize their research, including sections of questions on research topic selection, collaboration, and funding. The next section was designed to specifically elicit information about their scientific data repositories use and RDM practices and data deposit. In this section, participants were asked to narratively tell the story of their experiences with interacting with scientific data repositories, which all participants had used, designed, or contributed to. The main focus was on the scientists' experiences with

institutionally created and maintained scientific data repositories such as GenBank, Gene Expression Omnibus but also included online resources that the community designed and maintained, such as FlyBase (https://flybase.org/), Mouse Genome Informatics (MGI), and Saccharomyces Genome Database (i.e., Yeast Genome Database). A funneling approach was used to enable these narratives to emerge from the participants' description of how their lab came to be organized, as well as the participants' current behaviors, beliefs, perceptions, and attitudes about RDM and data deposit. Here, the 'funneling approach' refers to the technique of beginning with a broad set of questions (e.g., related to the background of the scientist) and gradually narrowing the questions to target participants' experiences with data deposit (e.g., uploading a dataset to FlyBase).

Most of the interviews (n = 9/12) were conducted in the office of the faculty participant (1 interview was conducted via telephone). Meeting in-person was the preferred mode, to ensure comfort and improve the participant recall about their daily work. Face-to-face was also advantageous to reduce disruption and observe the lab environment *in situ.* Photographs of the bulletin boards, conference spaces, and display cases were captured and fieldnotes were taken to document the department layout, facility size, interdisciplinary of the space, and other features of the laboratory space relevant to RDM and/or data deposit practices.

The average length of an interview was 76 minutes. For confidentiality, individuals were assigned a unique ID (P1-12) in chronological order of the date of the interview. Sensitive information was removed from the transcripts (e.g., names of persons mentioned, discipline and university-specific information that may disclose the person's identity). The audio files of the interviews were transcribed using the semi-automated software *Rev.com* and a verification of

transcription accuracy was executed by listening to each interview while following along with the textual transcript and correcting any errors while anonymizing the data.

The primary affiliations of the recruited participants included four universities, public and private R1 research U.S. universities (**Table 8**). Participants' disciplines were all in molecular biology and genomics, but spanned a range of sub-disciplines and specializations, including biochemistry, genetics, neurodevelopment and neuroscience, molecular physiology, psychiatry, and the behavioral sciences. However, participants differed in how they allocated their effort. In general, all interviewees had responsibilities of teaching, service, and research, though in different ratios. Nonetheless, the R1 status of all the institutions made research the participants' primary responsibility. Likewise, participants had different levels of experience, with the least amount of experience represented by the 1st year tenure e-track assistant professors and the most experienced as the faculty who were also department chairs and endowed professors. Despite these demographic and experiential differences, the interviews illustrate the richness of the career lifecycle and provide a generational perspective and more breadth of experiences to the shifting contexts of RDM and data deposit practices.

Three criteria identified participant research as a 'data-intensive': interviewee self-report, documentation (physical and virtual), and observations of lab environment (physical and virtual). First, faculty who explicitly described their research programs as 'data-intensive' using terms like big data, computational biology, or "data-driven modeling" were identified as data-intensive. Second, documents were located through online search such as from faculty professional websites or scientific publications. The documents often contained similar language of self-reports, such as dependencies of the research on frequent interactions with research data, computational artifacts, data workflows, and inscriptions of knowledge infrastructures related to

collection, (re)use, manipulation, and analysis of data. Third, the data-intensive nature of the research was additionally verified by scientists' descriptions of the nature of their work during lab tours, pointing out equipment as "big data" machines used for data-driven methodology, use of their computer for advanced statistical modeling of data and data mining activities, and the employment of machine learning techniques.

Data collection and analysis was conducted in multiple stages, beginning with online documentation search and synthesis about the faculty's research, fieldnotes taken on site, and interviewer reflections conducted prior to and after interviews. Next, the research fieldnotes, online documents, and transcripts were grouped by interviewee.

Content analysis proceeded by identifying themes and sub-themes, concepts, and patterns in the participant responses and analyzed in chronological order of interview. The codes were analyzed for relationships, and the concepts and topics were discussed, iteratively, with 3 rounds of axial coding (Elo & Kyngäs, 2008b). An inductive approach informed the three rounds of analysis, drawing from grounded theory (Strauss & Corbin, 1994). Grounded theory is premised on iterative coding of emergent themes, topics, and concepts.

The first round of coding involved re-listening to the audio files while simultaneously analyzing transcripts to identify codes, themes, topics, and relationships present among the responses. In this first round, the tool Microsoft Word was used to facilitate a less-structured coding environment which enabled the necessary function of surfacing emergent, inductive themes and open ended memo-ing (Charmaz, 2006). The second stage proceeded with data analysis using QSR NVivo 12 to provide a more structured approach to qualitative content analysis. Specifically, the NVivo 12 software (professional version) contains coding tools for inductive and deductive analysis, with features enable noting relationships among participant

responses (Elo & Kyngäs, 2008b). In the third round, a review and revisions of the codes were made. Where appropriate, the codes, themes, and memos were revised such as re-naming a code upon reassessing of the content, collapsing multiple code names into a single code, and/or splitting a code into additional codes if this parsing increased conceptual and semantic accuracy. From the coded transcripts, the categories were then generalized into supra-categories and subcategories, described in the section below (see *Findings: Categories and Codes*). The supra-categories and sub-categories and themes found were usually directly related to the interview questions (e.g., *factors contributing to the formation of interdisciplinary collaboration*). However, the codes and themes also included emergent content (e.g., *role of funding initiatives promoting data-intensive science in the visibility of inter-university collaborator*).

4.3 Findings: Categories and Codes

This section reports a synthesis of the exploratory study results. The findings are organized by the research questions of this study. First, the participants' descriptions of the data practices are presented and discussed. Next, the social, cultural, and organizational factors that contribute to how faculty make decisions related to RDM are related. Throughout this section, there are brief descriptions of the interviewees' insights related to the major institutional features that have shaped U.S. academic data intensive research environments. These institutional conditions described by the participants are salient for the research questions and to inform the conceptual framework that emerged from the categories and themes identified in interviews. In each section, the codes and constructs from the inductive analysis are reported.

The exploratory study was particularly important to elucidate the activities and artifacts that constitute faculty data practices. With a special sensitivity toward the formal support of these practices, eight categories (prospective theoretical constructs) were identified to represent

the emergent themes present in interviews: a) Administrating and Managing, b) Maintenance and Repair, c) Collaborating and Relational Labor, d) Archiving and Documentation, e) Socialization into Data Culture, f) Data Articulations, g) Authoring and Publishing Activities, and h) Risk and Uncertainty Management.

- *Administrating and Managing*: included the activities which participants mentioned that related to the bureaucratic organization and management of the lab group, especially as it directly related to data tasks, duties, and responsibilities. These activities include hiring computational talent and workflow management. In short, it refers to the overseeing and guiding of research group for core tasks related to managing the financial, hiring, mission and vision, and other lab affairs.
- *Maintenance and Repair*: the upkeep of equipment, materials, and skill base of the lab or research group. The training of student is considered a type of maintenance work, as well as feeding the model organisms and 'keeping up' with the pace of daily data-intensive science tasks and activities.
- *Collaborating and Relational Labor*: includes the activities – behaviors, emotions, and cognition (e.g., planning) – participants discuss that are related to cooperatively working with individuals both inside and outside the department. The interviewees experiences of interacting with others to form, maintain, and break off relationships is included.
- *Archiving and Documentation:* this category indicates the actions involved in preserving data and information, as well as making data, procedures, or results findable and meaningful through records-making and record management.
- *Data Culture*: refers to the process of teaching, learning, training, and becoming part of a disciplinary or laboratory systems of normative thought and behavior. This category

represents data culture at multiple levels, including the data culture of the laboratory, the professional data norms of a discipline, and the wider scientific environment to create a culture of ethics, rules, and norms around data practices. Debating validity threats through traditional channels such as the peer review process, informal and formal discourse and debates, and via professional associations.

- *Data Articulations*: represents the instances when participants mentioned having to formulate and execute tasks to connect the well-articulated overarching goals and milestones of a project by doing tasks that "makes the [data] work *work*"[6].
- *Authoring and Publishing Activities*: corresponds with the cycle of preparation, submission, and revision of formal and semi-formal knowledge products, such as publications, datasets, and preprints. This category includes interviewee accounts of activities such as authoring, telling a compelling story, identifying an audience, and submitting datasets. As such, this category which is not limited to the writing of papers.
- *Risk and Uncertainty Management*: refers to the strategies for avoiding and/or dealing with potential harms – e.g., staking intellectual territory, avoiding getting scooped, data errors throughout the research cycle and across areas related to data practices.

Within these categories, the data analysis surfaced 28 codes across the interviews. The sub-themes within each category are ordered in a manner that begins with the findings most closely related to the research questions and goals; however, an exception is the category of *Risk and Uncertainty Management*, in part because it is an emergent category (i.e., it was an issue not explicitly designed into the interview questions). The categories overlap and are related to

[6] The term *articulation* is used in the sense which Suchman (1996) defined the construct (see also Chapter 2 for literature on the concept).

eachother, cross-cutting codes and themes to form other constructs. However, given the exploratory stage of the analysis it is premature to draw relationships between the categories.

4.4 Discussion of Findings

In this section, select findings related to data management and deposit practices are articulated in detail with supporting quotes from the interview data analysis, along with the supporting literature. To scope the findings, the initial 8 categories and 28 codes were narrowed to focus on categories most relevant to the focus of the study: data deposit practices. Three categories were selected. I reviewed the themes from the original 8 categories and referenced the literature defining the purview of the activities that can be classified as data management related directly to data deposit practices. Here, I drew from literature on the data management "lifecycle model" (Carlson, 2014) and "data journeys" (Bates et al., 2016; Leonelli, 2016).

The resultant categories were *Maintenance and Repair, Archiving and Documentation,* and *Data Articulations.* The purpose of focusing on these three is that these are categories relevant to the focus of the study on data management and deposit practices. The focus of this study is on activities most directly on 'the path' to data deposit. For example, the category *Risk Management* was excluded. Granted, the *Risk Management* category includes mitigating data errors – an important antecedent to deposit. Nonetheless, the category was out of scope because it related more to organizational issues posing risks to the scientists rather than direct impacts on precarity during the daily data deposit practices (e.g., the "avoiding getting scooped" and "securing funding" were more general threats to the continuity of the project. Successful data deposit depended upon them but at an organizational level rather than the everyday data practices to prepare data for deposit.

4.4.1 Maintenance and Repair

Maintenance and Repair are the activities involved in 'keeping up' with the pace of daily data-intensive science tasks and project outcomes. They involve sustaining the lab by maintaining supplies, re-calibrating technical equipment, and updating the skill base of the research group. Here, training student to maintain the level of competency and renewing professional relationships to sustain the social capital of the lab are types of maintenance work. This work is distributed among lab members from the lead faculty researcher to the students but falls heavily to the technician. Interviewees described the vital role of the technician in their lab's maintenance and training of newcomers. In this category, four distinct but overlapping categories of work emerged: Infrastructural upkeep and repair, orienting newcomers, training students, and data 'community service.'

Infrastructural upkeep and repair: In daily lab practice, work is involved in maintaining the material upkeep of the lab (Denis et al., 2016). The technicians and student lab members often take responsibility for or are tasked with ensuring a clean and well-stocked work environment. The work is an exercise in training, such as when junior lab members are assigned the work of preparing fly food or cleaning animal cages. To the extent that data practices are activities that related to the production of meaningful scientific representations (Borgman, 2015), that is the "alleged evidence" (Edwards et al., 2007) for constructing facts, upkeep data practices include preparing of samples from which data is collected, ensuring that equipment does not produce errors because of contaminated conditions, and monitoring computational processes for technical 'glitches.' Interviewees highlighted the role of technicians in keeping the lab running through the maintenance infrastructural upkeep. As P5 describes, the technician is essential across all stages of the lifecycle, from re-stocking to analyzing experiments:

> *Yeah, so the technician, the role of the technician is to ensure the daily functioning of the lab. I guess routine tasks that are essential for making – for example ordering of supplies, ensuring stocks of particular chemicals and other commonly used reagents in the laboratory are kept in maintenance. So basically daily, the task of the technician is vital to the lab even though they perform the, I guess the mundane parts of a lab, but they ensure that the laboratory is functioning…. But they also perform experiments too, as well, which that is more directed by me. I provide instructions on what experiments to perform. So the majority of their tasks are to make sure the lab is running, but also, they do perform scientific experiments and they do contribute toward discoveries in our lab. So that's the role of the technician. (P5)*

Technicians or lab managers orient the junior students. In contrast to P5, an Assistant Professor, whose technician not only does maintenance but makes contributions "toward discoveries in our lab," P7, a tenured Professor has a clearer division of labor between the "actual researchers" and those who perform infrastructural upkeep and repair work. In P7's words:

> *We do have a lab manager. Who in theory does everything else to take the burden off of the actual researchers. We have a lab manager and then we have undergrads who make fly food and clean the lab and get rid of bottles and things like that…. Also, we have fly stocks. These are Drosophila, either species or genetic strains that have to be maintained and so the lab manager does that… they supervise the undergrads to do the dirty work or mind the flight kitchen, clean bottles and things like that. (P7)*

Orienting newcomers: As new students, postdocs, or faculty enter the lab group they often require assistance with becoming familiar with the processes and laboratory practices. Orienting newcomers is the effort of ensuring that new lab members are 'up to speed' in required knowledge, both technical and social. Lab orientation activities range from directing them on how to get ethical certification (e.g., CITI training) and explaining how lab data management is delegated to providing background literature and explaining lab norms. As P8 explains:

> *Now, the lab technician I have now oversees everything. And is like my eyes in the lab… where it used to be more that the lab technician would kind of worked for me and we worked together. But now I don't have time to go in the lab, so much. So she kind of oversees everything… That includes the training on the microscope, weekly lab meetings, she helps to schedule that…when each new person comes in the lab, they need to go through all this different training. The IRB with the Institutional Animal Care and Use Committee (IACUC). She also oversees our mouse colony... Like we have people that actually take care of them, but she oversees what we're doing with the mice….because basically we have to get newborn mice so we have to mate males and females all the time, so she's overseeing…especially when people first start in the lab and giving them things to do, before they know kind of what they do. (P8)*

In larger labs with high turnover, faculty tend to build more controlled processes and hierarchical training mechanisms for orienting newcomers. In mid-sized and large labs, orienting newcomers is done by the technician or designated lab manager or as a training for more senior graduate students, as P5 describes:

> *In my particular lab, I provide them with the opportunity, just like the post doc to mentor undergraduates. Which is a great opportunity for them to also become a leader and also to have someone help them simultaneously training the next generation of scientists. So we have a very vertical mentoring type of system where my graduate students mentor, typically every semester, one to two undergraduates, sometimes three, simultaneously in the laboratory. So basically, they're almost running their own little laboratories as well. I run a laboratory of seven personnel and then each one of them has an undergraduate that they're mentoring as well. So we have like many levels of labs in my lab." (P5)*

Similarly, P6 delegates the orientation of newcomers to graduate students:

> *My two graduate students…have, I call it like a little army of undergraduates who are assisting them with their projects. It's worked out very well in terms of the undergraduates get training in advanced techniques and otherwise…they get exposure to world class research, while simultaneously the graduate receives training on how to describe a technique or describe a particular concept to someone, which is also great training as well. (P6)*

The hierarchical organization of orientation is also evident in P8's lab management plan:

> *An undergraduate comes in as a sophomore, then there are junior, senior. So then they'll help train the newer ones coming in. So we try to all help train each other. (P8)*

Advising and training students: Although related to orienting newcomers, advising and training students is a distinct role in data-intensive science. Most commonly, the faculty assume an advising role. The advisor-advisee relationship can often be formalized, where faculty are committed to guiding students through a degree program. However, the advising relationship can be informal, such that the faculty acts as a mentor (Armstrong et al., 2002). In either case, faculty often work collaboratively with students on research projects as guides to both the science and the academic profession, what interviewees described as "developing the next generation of scientists" (P7). As P5 explains:

> *How it usually goes is that the PhD project in biology, essentially, it's preparation for them in their future of preparing for writing a grant and becoming independent scientists. So each one of them has a concrete set of aims, scientific aims that are all centered around a particular hypothesis that day that they come up with in consultation with me. And they defend that particular proposal to many and then after they successfully defend their proposal, then they, for the next four to five years, then perform experiments and other research to test their hypothesis and other particular aspects of their project. During this process, they lay out their vision for what type of project they want to do. (P5)*

Advising, training, and providing career advice to students was reported by P7:

> *Our students have their own projects. Typically, when they finish, they do the paper, they lead the writing. I have to, you know, advise them how to structure paper and help them to edit paper, find out the problem, how to write. It goes back and forth for many, many rounds. It's very time-consuming, because you don't*

> *want to do everything for them, otherwise they won't get to learn. That's a major headache in the lab. (P7)*

Faculty train graduate students as well as postdoctoral researchers. P5's postdoc joined the lab to get trained in new methods and techniques: "His role was, I guess it was to receive further training in whatever field that he was interested in. In his particular case, he was biology, biochemistry and so he joined my lab to further extend his training." The postdoc would also assist with planning scientific projects:

> *Of course, he had a voice in determining what type of project – in consultation with me – what type of project he wanted to work on in terms of day to day, week to week goals and planning, he also planned a lot of the vision for what his eventual aims were." (P5)*

Sometimes, training students involves learning professional best practices, such as the norms of communication in science and how to productively manage her time:

> *Because the students sometimes don't know any better, right? I mean, that comes with training students, and training them what works and what doesn't. That depends also on the personality of students. I've had very creative students that have had lots of ideas and this is a lucky thing to have, right? I could select, of all of the ideas she was throwing on the table, which ones were more valuable, and still keep her focused, and just saying, "I know that you're productive, but you cannot do all of these things. You just have to choose what you want to do." Right? That's the best-case scenario. (P6)*

Data 'community service': Maintaining the sociotechnical infrastructure of community databases emerged as an important theme in the interviews. Interviewees discussed how they contribute to open scientific research databases not only as a formal requirement of some publishing venues, but also as a form of "community service" to maintain and curate the database resources. Smaller, less centralized databases such as FlyBase and the Yeast Genome DataBase (SGD) are looked after by volunteers who maintain the community resources. The need to support model organism databases and scientists attitudes toward 'pitching in' evidences a logic of care for both the community and the continuity of their profession's data infrastructures (Edwards et al., 2007). As P7 explains, the sequence database for *drosophila melanogaster,* called FlyBase, is the product of this logic of care:

> *There are people who are actually going through the literature. Remember, there's fifteen thousand genes in drosophila... We don't know functions about a lot of genes, but people are actually writing this manually. These are manual curations. There's a team of people who've curated this information. But this just goes on and on forever. This is all information that has been curated about this one gene… There's a team of people integrating information from publications in real time. Now obviously, there's more information out there than you can ever make ... than you can keep up with to populate this thing. But also you can add personal communications to FlyBase. So these are non-published. So non-published notes and personal communications of that are of relevance to this gene. This is all been done by somebody. There's a team of people who are involved in this. So this is just an example of a research community where there is an incredible effort to share and collaborate and make useful information available to everybody. (P7)*

Yet, concern for the quality of the data curation and metadata drives the efforts of scientists to try to support the upkeep of the community data infrastructures:

> *For the first time now they're starting to ask for contributions. FlyBase is this long-standing…repository of all information about* Drosophila. *And you can see here, they want people to actually sign up and pay for it, because of course the NSF is cutting their budget. But this is an amazing resource to the community. It has all this data….But they say here, look. It actually NIH. "The NHGRI is significantly reducing the funding for FlyBase by 15%. With these cuts we're not going to be able to deliver high quality essential curation and tools." So they're trying to raise money from the community because they realize the curation standards will decrease. (P7)*

Although the data curation work can be through the unpaid efforts of faculty, much of data curation is a behind-the-scenes job which is completed by staff scientists at institutional repositories such as GenBank. How precarious a database and its data are varies. As P9 describes, the database is maintained only if there is funding, long-term planning, and differs between industry and academia:

> *Now it's self-sufficient…. [The database] is in Seattle, and they have their own scientists. They hire their own scientists, but their original mandate was to look at the expression of all the genes in the brain. They do things in higher throughput that most of us academic scientists wouldn't do because you can't get that funded. You can't publish that really, but it's data everybody can use, so they do all this and now they're doing in human as well. They're doing OS expression data and genomics and things looking at single cell expression profiles and data, and all of it gets updated immediately into this website. (P9)*

The interviewees derived value from the community database resources, along with their larger more centralized and institutionalized counterparts such as the Gene Expression Omnibus (GEO) and GenBank. Participants (P1, P3, P7) explained the linking system between the larger repositories and the niche community repositories in terms that suggested they supported the

ideals of open science, even as there could be extra work in contributing to the maintenance of the data community infrastructures.

4.4.2 Archiving and Documentation

The work of documentation and archiving is recognized by a generation of literature as invisible, often gendered, labor (Suchman, 1994). Studies of the invisible labor behind smoothly running digital systems have included examinations of the preservation work that librarians do to enable the "dream of the automated archive" (Paisley, 1968; Star & Strauss, 1999), the "data labors" involved in the archival activities of metadata information professionals development for open scientific research data repositories (Nadim, 2016), and the "ghost work" of gig workers that curate online content (Gray & Suri, 2019). Interviewees discussed a range of crucial activities of documentation for their research lab and the community at large (P1, P2, P6, P8). Of key importance are keeping consistent and up-to-date records, the development of a sustainable lab or community database lab, and the use of dedicated lab notebooks as part of labs' data workflows. Four sub-categories constitute this category: planning for data management, lab notebook keeping, computational compatibility, and navigating data standards.

Data management: Research data management has been a global issue and central concern for scientists in data-intensive research (Atkins, 2003). The work of data management from system design and data curation to organization for searchability and longevity has often been left up to individual researchers and disconnected institutions (Qin, 2013). Developing competencies for researchers only developed within the last decade (Crowston & Qin, 2011), and a substantial workload falls to faculty to manage data. The pillars of data management described by Qin (2013) are institutionalization, standards, and infrastructure. Throughout the research lifecycle, the academic researchers interviewed here discuss the importance of data management. First, data management at the local level is a high priority. Interviewees describe a mix of digital and

analog formats through which the invisible work of ensuring the data is secure, accessible, and updated are mediated.

The data management practices in the lab reflect a *bricolage* of digital and physical materials – a term adopted by STS scholars (Ciborra, 1992; Nardi & O'Day, 1999; Sawyer et al., 2011; Star & Griesemer, 1989) to refer to the ongoing construction of information systems by bringing together a diverse array of available items and ways of working. In short, the data management arrangements of in interviewee labs reflected not only the bottom-up, improvisational character of how data artifacts are organized and curated but also revealed how data practices are contingent on the division of labor and knowledge(s) imbricated in the lab group and space. P6 describes the bricolage of her lab's data management:

> *We had a server, where we had protocols and things like that, but at the same time, I have to say, I'm old-style. I feel that all these things that are digital are excellent if you have a lab manager that is pushing everyone to comply and do it. For instance, we have protocol books like binders that are protocol books, and then we tried to make all that digital. We have a departmental server where we all, just with our Net IDs and credentials from the university, we all can access it and we can just put it in our computer and we see it like a folder in our computer. (P6)*

This quote overlaps with the sub-category in administrating and maintenance work, that of protocol adherence and enforcement. The overlap suggests a close relationship of documentation with data management. As P2 describes, her lab keeps records with a forward-looking vision of lab turnover and the goal of lab data longevity to avoid a common problem, unrecoverable data. A data management best practice is to keep the original file and "let people take a copy with anything they want," such as the experimental protocol, for instance, "how you are doing the subcellular fractionation" (P2).

Lab notebook keeping: Documentation and archival work is required for records be kept over the course of multiple generations of students, enabling the data's "mobility" and longevity (Latour, 1987; Stöckelová, 2012). Lab notebooks are a staple artifact for documenting scientific

processes and practices, as some of the classic anthropological sociological studies of scientists and engineers, e.g., in the biological, physical, and ecological sciences have reported (Latour & Woolgar, 2013; Shankar, 2004).

While digital notebooks and repositories such as the computational notebook Jupyter and the code repo GitHub are a signpost of open data-intensive science, interviewees keep results protocols within the lab, both digitally and physically for a variety of reasons related to lab culture or computational (il)literacy (Kluyver et al., 2016; Randles et al., 2017). For instance, P2, Professor and Chair of Biochemistry and Molecular Biology at a public research university, explains her motivations for notebook use, rules and best practices, uncertainties related to digital or analog medium of notebooks, and the data management culture related to lab notebooks. In P2's words:

> *We have a lab notebook that's chronological describing what you do each day, right. Inserting data as you can. For the different types of experiments you need to download [the data] from the microscope. We want to keep that by date too. So you have sort of pointers to the data in a lab notebook. So usually they put a file name. The really traditional thing is this bound thing and everything's in there…. But a fully electronic notebook scares me because I also feel like I am old enough that I've seen media, I've seen you can't read all things always. Whereas my lab books from 25 years ago are still there literally right there. This is an ongoing issue. I have things on this that I no longer have the ability to read. And if you think of your data, not just in the, I think it's like three to seven years or something, but you know for really long term there's nothing like the written documents. So this is a big, in our field, that's a big thing right now. (P2)*

Lab notebook keeping has a documentation function, but other functions as well. For instance, P5 and P6 have digitized lab notebooks to perform validity checks and trust. Slack and electronic communication function to keep records and document informally. Archiving and documentation also includes sharing the notebook and maintaining best practices according to the lab rules and norms of lab notebook use and information entry.

Planning for computational compatibility: A concern for interviewees is anticipating and planning for computational compatibility to avoid unreadable data, outmoded formats, and inaccessible documents. Related to data management and lab notebook keeping, this category

was an emergent theme in interviewee conversations of addressing the collision between file formats suited for cutting-edge research and file formats suited for archiving and preservation. Interviewees planned for backward and forward (re)-engineering to integrate old formats into new ones and preparing new ones to have longevity. The more senior interviewees (e.g., the professors with the longest scientific careers in this participant sample: P2, P3, P4, P7, P8) who had been consistently working with data-intensive research highlighted the issues of computational compatibility:

> *I just am worried about…I think people haven't thought through that. I can't even read a word file from 10 years ago…These Lab notebook programs are quite specific, that company goes out of business….Are you going to be able to read it? That's my primary worry. There could be security worries too, but like I told you, I'm not as worried about that. We're really paranoid about backing everything up too, but that would be a worry that people forget to back it up and it's gone. One of the computers, which was old, just crashed and they're like, "Oh gosh, Oh my gosh." But nope, there was an internal hard drive. It had been set up and you're like, yeah, so, but stuff happens and the idea that it would be gone, gone is really scary. (P2)*

The worries about compatibility have been assuaged by backing up information but also the use of more universal formats and technologies. Yet, just because a tool or format is widely used , that is, among many people and organizations, tools such as Microsoft (MS) Excel, does not guarantee the long-term preservation of the data stored in the tool. For example, data files stored in Excel can become inaccessible due to deprecated software. Likewise, computational technologies that are new to the lab are a cause for wariness and a need for more vigilant compatibility checks, as P2 explains about raid arrays:

> *We found out a microscope was storing via raid array, which he hadn't realized. And so a raid array saves kind of, it has to be very fast because it splits the data. And that data was just basically unrecoverable. I mean we could have [recovered the data] for hundreds of dollars and it's like, what do we really have there? Nothing we can't do without [a] problem. But it was just– I just hadn't realized it was savedthat way, you know? (P2)*

As P2 highlights, the need to plan for compatibility is an ongoing data practice that can often go unrecognized as an important data management step within individual faculty laboratories, and at a higher infrastructure and institutional level of research data management.

Navigating data standards: Data standards have created opportunities to link data across repositories and to create more stable, accessible, and interoperable systems for intractable research data management issues (Wilkinson et al., 2016; Zeng, 2008). Academic research faculty benefit from the standardization but also have added tasks. For example, faculty must learn the new data standards, adapt legacy technoscientific artifacts and practices to new data standards (Stöckelová, 2012), and interpret standardized data whose idiosyncrasies have been obscured by standardizing technologies.

On the other hand, the added work of navigating standards has benefits for scientific output. As P3 and P4 express, standardization has facilitated finding and retrieving data, and is a feature of model organism data, as a result of the reaching a critical mass of data production and collective needs of a specialized community:

> *Right now the yeast genome has been sequenced for 15 or going for 20 years. So new sequences are no longer an issue, right? So it's screened now. So I sent them a couple of, of data for our screens. I asked them what they want and I forget what they want. They're willing to take like Excel spreadsheets where I had indicated strength of phenotypes. So they were willing to work with that. That is a wonderful example of a genome repository. And there's certainly like for, for people to do structural biology, they have to submit to PDB (protein data base) with a release time of a certain amount in order to have the paper be published. (P3)*

The institutionally implemented metadata, formatting policies, and other standardization features are encoded in research data technologies, such as open data repositories, but also permeate the data artifacts and practices of specific disciplinary communities (e.g., Fly Base, Saccharomyces Genome Database) and of individual labs. The ease of use for scientists is enabled in part because it has existed for decades and now there is a system, processes, workflows, and greater compatibility with technologies such as those in commercial file formats (e.g., Excel in Saccharomyces). The computational literacy required to use the database is a lower bar because of this standardization, that is, there is no need to learn many esoteric file

formats (*Cf* Star, 1999 where the technical literacy requires of using Worm Base posed a number of learning and literacy challenges for scientist-users).

Standard formats are developed by a community of practice. Over time, database standards can shape practices and the configuration of scientific work (Bowker, 2000), what Robert O'Hara (1992) referred to as the "grooving effects" of standardized formats in tree databases. It is out of this study's scope to explicate the origins of sequence databases, and the work to develop them, how people learn to use them, and how they become reinforced in practice. Yet it is crucial to note the standard formats for DNA data were developed in a particular historic moment. For example, the standard formats P4 describes as "universally accepted" and that "everybody uses" were not inevitable. As in the case of tree chronology (O'Hara, 1992), data formatting standards according to a path of "selective simplification" in which the data more easily represented according to existing schema tended to be deposited. For example, the data is deposited if it is easily represented in the required formats and current schema. There are critical implications for what data is deposited. As a result of the 'path-dependency' of the development standards – in other words, the historical contingency of what standards are created – a set of standard formats, "data structures and information retrieval models are set up so that a particular, skewed view of the world can be easily represented" (Bowker, 2000, p. 661).

In sum, documentation and archiving involve preserving data and information, as well as making data, procedures, or results findable and meaningful through records-making and record management. Data standards has facilitated some aspects of work, while adding more, creating additional invisible labor for the 'behinds the scenes' personnel of standardization work.

4.4.3 Data Articulations

Data Articulations represents the instances when participants mentioned having to formulate and execute tasks to connect the overarching goals and milestones of a project by doing tasks that "makes the [data] work *work*" (Suchman, 1994). The interviewees described their need to deal with unexpected situations and re-engineer workflows and activities on a regular basis. The four sub-themes within the category of *Data Articulations* include data bricolage, planning data-intensive workflows, and revisions and retractions.

Data bricolage. Data bricolage refers to the work of piecing together the parts of the infrastructure into a coherent whole. This is creative work that requires bringing together seemingly unrelated projects, skillsets, equipment, and data to achieve project results. In the case of P9, assembling the available tools – at a low cost – is crucial for her management of the lab. P9 had to be, as described an instance of creating her own equipment in house from a glue gun and GoPro:

> *The basic scope I have now is already $100,000 but we could add capabilities to that for example and going through another thing we do is test behavior in mice… Some of these setups you wouldn't believe what they charge in science. But it seems like a plastic box but it's about $12,000. So, we've made our own out of corrugated plastic with glue guns…. And we use a GoPro. We do a little more rigged up, but I would put equipment into keeping so that my lab can do things like that where it would be a little more flexibility like a trade-off between wasting money and just saving the time where we could purchase things like that going forward. (P9)*

In the context of knowledge work, *bricolage* is a term that refers to the practices which individuals engage in to bring together disparate entities to assemble a working arrangement (Erickson & Sawyer, 2019). Here, P9 assembled a cheaper version of a device that costs $12,000 with items including glue guns and a GoPro. The artful configurations (Vertesi, 2014) are one type of data articulation, bridging gaps where otherwise their goals would be cost prohibitive.

Planning data-intensive workflows: data articulation work also involves creating a moving schedule around the lab's research agenda, with timelines with ever-evolving goals. Interviewees

discussed how dropping or reviving projects is part of the planning process, and like data bricolage requires agility depending on how the project is developing. As P7, a tenured professor described data-intensive research necessarily entails rethinking and reconstructing systems to achieve your initial plan:

> *You learn a lot doing research. Your focus is always changing as you learn about what you're doing. Sometimes you realize that the way you started off initially was misguided or naïve. The data is not publishable the way we wanted it to be. There's a lot of two steps forward, one step backward. There's a lot of rethinking…and redoing it. (P7)*

The scheduling and timeline setting such as for important deadlines and benchmarks, a type of articulation work that involves keeping the team up to date and keeping the momentum and project "on track" (P5). Often, the keeping of the timeline is done by a lab manager and to manage the smaller "moving parts" of the project as deadlines approach. However, the goal setting on a larger scale is the role of the faculty. As P5 illustrates, there is articulation work in setting timelines because of the constant updating intrinsic to uncertain projects, and are set "day to day, week to week" and work that is distributed across students, technicians the PI:

> *The primary role for determining timeline, our goals, our scientific goals, such as meeting, we have particular, specific aims that we aim for in terms of trying to, scientific aims that we go for. Those are primarily set by myself or the graduate students in their individual projects. The technician is more so for the responsibilities…running the lab meetings. But for setting timelines and goals, that is less so. They're more on a day to day and week to week where I provide much of the instructions for what are we doing this week. They provide feedback on their schedules on whether what is feasible within that week, but I primarily set the schedule of the week and the months. (P5)*

Revisions and retractions: related to data articulations, the unexpected modifications to an analysis, or a retraction of an error, is unexpected work that requires additional labor by the research lab. Revisions are most commonly called for by a publishing editor and may include revisions to the data documentation, or an addition to or revision of supplemental materials or submitted data. Often, data revisions involve a request from journal editor to add additional samples, run more data analysis, or design alternate experiments. For instance, P6, an Associate Professor in Genetics, who is also a mother and expressed how she cannot spend more time than

she allots for herself at the lab, explains how additional experiments can be requested and take more time than expected:

> *If [the paper] gets rejected you have to submit to a new journal. If they ask you for experiments, put a month and a half, two months, three months of doing experiments, right? So there you have, now, like the two months, plus the three months of doing experiments and you don't even have the paper accepted, right? So then you resubmit and it's another month or a month and a half, right? So you're already more than half a year plus then it gets accepted on the second route. So eight months is the minimum that you get a paper accepted and in eight months a lot of things can happen. So we didn't do that many experiments but the few experiments I had to do them, so I had to find the time to do it and I had other priorities that's jumped into my table that I had to take care of. (P6)*

Moreover, the reviewers may "complain about text, ask for more experiments… It can be from more data, better data, rewriting the conclusions to match the data" (P3):

> *[Papers] have to be revised, you have to have more data, you have to fix things, and then they get accepted or their revisions. So the more as we go along, reviewers want more and more and more data.*

Data articulations help scientists to deposit their data and by create bridges between places where the work is underspecified. Three sub-categories of data articulations are discussed here: *data bricolage*, *planning data-intensive workflows*, and *revisions and retractions*. While not exhaustive, the three sub-categories serve to demonstrate the ways that scientists do not execute workflows in linearly rigid routines but develop workarounds and "artful" reconfigurations to manage and deposit their research data. Data bricolage shows how the scientists employ such artful reconfigurations to order competing priorities (e.g., by reducing costs of equipment to reallocate funds to personnel).

Planning data-intensive workflows is a strategy for creating structure for an otherwise underspecified work context, that is, the context of the PI-driven lab. Scientists can expect to get revisions and even retractions. But scientists cannot have precise expectations as to the details of the revisions and retractions, such as how many times they need to speak with the repository, or whether they need to generate new data. To address unexpected revisions, articulation work functions to pull together the labor and resources needed to finally meet the publisher or

repository requirements (e.g., delegate to a postdoctoral fellow the task of sending the negative data to the repository).

4.5 Contributions and Limitations

The limitations to the exploratory study are related to scope, population, sampling method, data analysis, and generalizability. The scope of the study was bounded to a focus on the data practices of U.S. research-active faculty in R1 universities in disciplines data-intensive molecular biology and genetics. The particularity of recruiting from a single discipline and the focus on data practices constrains the study findings to the faculty perspective and cannot be extrapolated to apply to discipline outside of data-intensive molecular biology researchers in R1 universities. Second, the population was limited to a small number of participants and were majority tenured male professors with successful labs. While the participants did include, four women, two pre-tenure individuals the findings are anecdotal and would require cross-disciplinary extension as well as greater representation of junior scholars, non-faculty science workers across gender, race, class, and sex. A more inclusive sampling allows for a wider generalizability of the visibility of R1 academic molecular biologists' data practices. Third, the data analysis was coded twice but lacked secondary annotation by another data coder to run inter-annotator agreement.

Notwithstanding these limitations, the findings of the study generated analytic categories to provide a foundation for the research framework for this dissertation study. Drawn from the interviews with active researchers, the experiences of U.S. academic faculty surfaced sociotechnical factors which was supported by and added detail to existing literature in data management practices in data-intensive science. In addition, this exploratory study assisted with contextualizing the longitudinal mesoscopic patterns of data authorship in cyberinfrastructure-enabled collaboration networks within the perceptions and attitudes of biosciences faculty.

4.6 Chapter Summary

This chapter provided an initial foundation for building an understanding of the data management and deposit practices in academic research labs. Premised on empirical findings of the data management practices in data-intensive molecular biology research, the chapter surfaced 8 analytic categories with 28 codes and corresponding theoretical constructs, highlighting the three categories targeting RDM and data deposit. In the next chapter, a conceptual framework is developed from the initial findings to be refined and extended in the study focusing on the genomics community to develop the finding of data “articulations” surfaced in this chapter.

CHAPTER 5
STUDY 2

5.1 Introduction

The field of genomics has fostered the growth of a robust cyberinfrastructure over the course of decades to make genomic data portable and ready for AI-enabled genomics. AI-enabled genomics depends on large volumes of datasets are training sites for machine learning. The genomic datasets that come into a database are therefore valuable inputs for drug discovery, etiology of diseases, and precision medicine. Cyberinfrastructure for genomics models datasets as inputs that enable outputs for society, adding institutional pressure on scientists to share their datasets.

The grand visions of CI-enabled outputs of science though elide the ways in which datasets get made and become inputs into the system, resulting in a gap between science policy goals of supporting scientists and promoting long-term research data sustainability. Grand visions of cyberinfrastructure in genomics to support AI-enabled genomics though hit up against the reality of producing and creating datasets for sharing. Prior studies that look at the relationship between data sharing practices of scientists to cyberinfrastructure show that first, that scientists rarely share their data. Second, datasets deposit is dependent on the goals of the scientist. Third, that disciplinary norms shape data practices of scientist and that data practices are very local. In general, the antecedent work that scientists engage to deposit to datasets to a cyberinfrastructure shape the kinds of datasets enter the broader cyberinfrastructure assemblages and are tapped for data-intensive genomics.

We take this one step further to argue that scientists who deposit their datasets into a data repository engage in data articulation work which encompasses the work to align the

epistemic, material, and ethical dimensions of data. Drawing on interviews with molecular biologists who are faculty at R1 universities, we ask what kinds of data work scientists do to make their data deposit *do-able*. We use Fujimura's (1987) lens on articulation framework, to show the alignment work that scientists engage in to make their data deposit 'do-able'.

Genomic scientists maneuver among the experimental level of data production, to the lab level of data organization, and the social world of data repositories weaving and reworking datasets to meet experimental thresholds and data repositories thresholds before datasets become ready for deposit. Scientist are making ongoing judgements about the validity of their datasets, the accuracy of their datasets. Internal thresholds showcase the kinds of alignment work that take place to make data acceptable to be deposited. Each type of threshold — material, epistemic, and ethical — serves as a winnowing check point through which data is made do-able for AI: made fit for deposit, corrected for discrepancies, brought up to preservation standards, prevented from committing ethical violations such as falsification or manipulation, etc. Scientists bear the burden of understanding their datasets and tailoring it to meet the needs of data repositories. In sum, our focus in this empirical study on articulation work draws attention to the kinds of alignment work necessary for datasets to become part of the training sets feeding AI.

5.2 Background

Genomics is a field of biology that focuses on the structure, function, mapping and editing processes of the genome through the interrogation of large volumes of DNA sequencing data with a combination of other data including, for example, clinical trial data. In the last few decades, genomics has grown a cyberinfrastructure to develop a pool of "big data" for AI genomics. By investing in advanced technologies, techniques, databases, and analytical methods for over several decades, as it stands genomics has a robust cyberinfrastructure for a data

pipeline to conduct AI genomics (Daugelaite et al., 2013). This mechanistic cyberinfrastructure provides clear scripts on how, where, and when to share data to build a pool of data for AI. As it stands, current cyberinfrastructure sets up the idea that datasets simply move between sources.

5.2.1 Standardizing Data Production and Institutionalizing Data Sharing in Genomics

At the turn of the 21st century, Western science entered a period of scientific research, commonly referred to as the fourth paradigm of scientific discovery, characterized by high levels of computation, collaboration, and data-intensive activity (Hey et al., 2009). Federal funding agencies have invested in this paradigmatic shift heavily, launching cyberinfrastructure (CI)-enabled initiatives (eScience in the U.K.) and an x-informatics capable of "revolutioniz[ing] science and engineering" (Atkins, 2003; Hine, 2006).

Motivated by the goal of asking new questions and addressing bigger, more complex problems by applying sophisticated data science techniques to expansive data repositories, scientists and funders alike have slowly evolved the nature of scientific work over the last two decades (Bozeman & Boardman, 2014a; Leonelli, 2014a; Schneider et al., 2019). This evolution can be seen in the rising scale and scope of 'big data collections' for aggregating and mining large datasets, a move that is underscored by the concomitant development of federal data management policies at the U.S. National Science Foundation (NSF) and National Institutes of Health (NIH) (Diekema et al., 2014; Sewerin, 2015; Van Tuyl et al., 2015).

The emergence of new institutional pressures in big science disciplines like genomics, are evidenced by data mandates that require scientists to deposit data in compliance with the federal policies, best practices, prominent tools, and social attitudes deemed most worthy of advancing long-term data stewardship (Crowston & Qin, 2011; Sands, 2017).

The field of genomics has built a digital pipeline for data by investing in technologies for producing, storing, and preserving DNA sequencing base pairs. The implementation of

cyberinfrastructure to standardize data production in genomics began decades ago in the early 1990s by federal investments to advance DNA technologies and DNA data standards. The Human Genome Project (HGP) is a notable event that established field-wide consensus to develop technologies for DNA sequencing and codify data standards and storage for DNA sequences through the formation of the Bermuda Principles (Jones et al., 2018; Marshall, 2001).

The HGP established the policies for rapid release of DNA sequences and made that data portable for reuse. Continued advancements in technology and genomic techniques such as DNA cloning, X-ray crystallography, DNA sequencing, DNA synthesis, amplification by the polymerase chain reaction (PCR), and transgenic animals rose because of grants and proposals to support the innovation in genomics (Gales and Mccormack, 2003; O'Driscoll, A., Daugelaite, J., & Sleator, 2013). New technologies and techniques began producing other kinds of large data.

Not surprisingly, genomics as a field built and supported databases to store, preserve and provide access to voluminous data for reuse. Databases and data repositories support the curation and preservation of large volumes of datasets being produced by genomics research through standards. DNA sequencing databases like GenBank, genome browsers, model organism databases, molecule-based databases, process-based databases, community databases are just a few of the kinds of data support systems that formed (Lathe III, Williams, Mangen, Karolchik, 2008). Databases and libraries rely on the nature of DNA sequencing dataset to inform their organization and preservation. DNA sequencing data consists of four-letter base pairs. Because some genomics data are represented as finite string in predictable patterns of characters, the design of the metadata annotation for the genomics databases were more straightforward than the design of data sharing might otherwise be (e.g., multimedia formats) (Bala & Gupta, 2010; Benson et al., 2017; Nadim, 2016).

Brought on by the recognition that data repositories will not fill themselves up with datasets, genomics as a field, through the development of CI, institutionalized data sharing as well. As put by several clinicians, ‘genomics has a robust culture of data sharing’ (Byrd et al., 2020). Data sharing in genomics emphasizes sharing to public and open databases. This culture is fostered by funding agencies, journal publishers and professional societies in genomics that have instantiated some combination of open data, data curation, data management, data release, and/ or data deposition policies. Being open and sharing data is a norm not an exception in genomics. This institutional environment bears weight on scientists and pressures them to contribute their datasets towards a cyberinfrastructure ripe for AI genomics research.

The larger body of funding agencies that scientists rely on for their funding have developed and instantiated data sharing policies tied to proposals and grants. The National Institutes of Health (NIH) is the premier funding agency for genomics research. The NIH instantiated a data sharing policy beginning in 2003. Since then, they have updated their policies including Genomic Data Sharing (GDS) policy in 2014 and the Dissemination of NIH-Funded Clinical Trial Information (Clinical Trials Policy). The National Science Foundation (NSF) is another federal agency that grants genomics research funds. The NSF requires a Data Management Plan (DMP) for all research proposals and grants. Funding agencies play a big role in shaping the data sharing practices of genomics researchers by placing mandates on data curation, data management, and data preservation with grants.

Publishers in genomics mandate data deposition with journal submissions. Elsevier journals Genomics Data and Genomics require data to be submitted to those journals. The journals Nature and Science where breakthrough genomics work is published have made requirements to store and share data. Moreover, professional societies maintain open data

policies and engage in workshops to promote data sharing. The Genetics Society of America is a scientific society composed of scientific researchers and educators in the field of genetics. In 2010 they committed to open data and in 2018, partnered with *Figshare* as a platform to share data. Corresponding journals of the societies have also mandated data sharing policies as well.

Library and archives organizations provide a wealth of resources for data sharing, curation, management, and preservation for genomics research. Organizations not only hold for preservation, but they are active sites for data consulting including the benefits of data sharing and guidelines to share data. A genetic database may provide services (e.g., metadata), products (e.g., software), and artifacts (e.g., datasets) structured to enable the search and discovery of "genes, gene products, variants, phenotypes… to enable users to retrieve genetic data, add genetic data and extract information from the data" (Durmaz et al., 2015). They curate nucleic acid datasets and provide products and services (Benson et al., 2017). Wider initiatives such as FAIR (Findable, Accessible, Interoperable, Reusable) is a standard for sharing data.

In sum, genomics has configured a robust cyberinfrastructure composed of technologies, policies, standards, and databases for scientists produce, store, curated and enable the flow of data across organizations. Over time, data sharing became institutionalized which placed constraints scientists to curate, manage, and/or deposit their datasets. This places a mechanistic valence on data flows in genomics specifically that datasets will enter a repository, be stored for the long-term, yet at the same time be ready and waiting for AI genomics. What gets lost in this mechanistic view of cyberinfrastructure is the human work it takes to brings datasets into a database or repository.

5.2.2 Scientific Data Practices

If cyberinfrastructure serves as a site for downstream inputs into AI genomics, then the data practices of scientists are the upstream sites that make possible a viable cyberinfrastructure to tap. The data practices of scientists are the human activities and practices that go into making a dataset, publishing a dataset, or depositing a dataset into a repository. Following the data practices of scientists provide a variety of opportunities to understand when scientists successfully deposit data – or fail to deposit data – including the practices that lead up to deposit. We can identify the evidentiary objects for AI genomics – i.e., datasets – that enter repositories by locating when, where, and how datasets get deposited.

However, what slips out of this purview to support and enable AI in genomics is the backend data work of scientists who create datasets that feed into the cyberinfrastructure and subsequently machine learning models. Understanding and creating a journey for pooled datasets through cyberinfrastructure as a marker for its value for AI genomics misses the point that datasets have interpretative flexibility which begin with the dataset creator (Leonelli, 2014a).

5.3 Methodology

The methodology was selected to examine the data practices of genomics faculty researchers to develop a theory to explain data deposit. We observed the substantial investments in cyberinfrastructure (CI)-enabled science and data repositories have led to a flurry of interest in effectively supporting genomics and AI. The topic is of concern to a range of stakeholders, from science policymakers and university research faculty to academic library professionals. Yet despite substantial federal investments, academic research in genomics proliferates with widespread data depositing work while lacking analytical framing for how data gets accumulated in repositories and the impacts on work practice and data quality.

This study examines the experiences, perceptions, and data practices of genomics faculty in R1 research-focused U.S. academic institutions (the "R1" in the Carnegie Classification). To examine issues of data depositing work, we posed the research question: *How do genomics faculty members deposit datasets into repositories?* (RQ1)

5.4 Data Collection

Semi-structured Interviews were conducted with research active faculty members (n = 18, 6 female, 4 untenured) within the research community in molecular biology with a focus on genomics and data-intensive methods (e.g., proteomics, metagenomics). Participants were selected for geographical proximity to control for state policy and facilitate site accessibility (Northeastern U.S.). Researchers from public and private R1 universities were recruited to provide diversity in teaching, research, and service expectations.

A non-probabilistic purposive sampling technique was used to select participants, a technique used to recruit participants with predefined characteristics that are relevant and informative for addressing the research questions (Oliver & Jupp, 2006). Small- to medium sized laboratory groups with a range of academic ranks, genders, and biological sub-disciplines.

The target population was identified first by the self-identification by scientists whose disciplinary expertise on their professional or department webpage indicates they are in the genetics and genomics community with a secondary criterion using the manual web search approach was how the department classified the faculty.

Table 9: Study 2 Participant Demographics

ID	Gender	Position/Title	Discipline
P1	Male	Professor	Biochemistry & Molecular Biology
P2	Female	Professor, with a leadership role at the department level.	Biochemistry & Molecular Biology
P3	Male	Professor	Biochemistry & Molecular Biology
P4	Male	Professor, with a leadership role at the university level.	Biochemistry & Biophysical Sciences
P5	Male	Assistant Professor	Biochemistry & Molecular Biology
P6	Female	Associate Professor	Molecular Biology & Genetics
P7	Male	Professor	Biochemistry & Molecular Biology
P8	Female	Professor, with a leadership role at the department level.	Molecular Biology & Genetics
P9	Female	Assistant Professor	Molecular Biology & Genetics
P10	Male	Professor	Genomics & Biochemistry
P11	Male	Professor	Molecular Biology & Genetics
P12	Male	Assistant Professor	Molecular Biology & Genetics
P13	Male	Associate Professor	Genomics & Microbiology
P14	Male	Associate Professor	Genomics & Microbiology
P15	Female	Associate Professor	Genomics & Microbiology
P16	Male	Associate Professor	Molecular Biology & Genetics
P17	Male	Professor	Genomics & Microbiology
P18	Female	Professor, with a leadership role at the department level	Molecular Biology & Genetics

Interview questions were designed to elicit experiences, attitudes, and behaviors around data collaboration and surface the data activities, perceptions about data visibility, and data cultures of the research group and discipline. The interview questions were focused on the workflows in publishing a paper and how data gets deposited to a research data repository. Questions were structured in a funneling approach, starting with broad questions about how Molecular Biology faculty organize their research, including sections of questions on research topic selection, collaboration, and funding. The next section was designed to specifically elicit

information about depositing data to scientific data repositories in the process of producing a publication. Participants were asked to narratively tell the story of their experiences with interacting with scientific data repositories, which all participants had used, designed, or contributed to. In our analysis we focus on the scientists' experiences with institutionally created and maintained scientific data repositories such as GenBank, Gene Expression Omnibus but also included online resources which the community designed and maintained, such as FlyBase (https://flybase.org/), Mouse Genome Informatics (MGI), and Saccharomyces Genome Database (i.e., Yeast Genome Database). A funneling approach was used to enable these narratives to emerge from the participants' description of how their lab came to be organized, as well as the participants' behaviors, beliefs, perceptions, and attitudes about collaboration and data practices.

Interviews were conducted in the office of the faculty participant (n = 13/18) or using video conference or telephone. Photographs of the bulletin boards, conference spaces, and display cases and field notes were taken to document the department layout, facility size, interdisciplinary of the space, and other features of the laboratory space relevant to collaboration and/or data practices. The average length of an interview was 76 minutes. The audio files of the interviews were transcribed using the semi-automated software Rev.com and a verification of transcription accuracy was executed by listening to each interview while following along with the textual transcript and correcting any errors while anonymizing and de-identifying the data.

5.5 Data Analysis

An inductive approach informed the three rounds of analysis, drawing from grounded theory (Corbin & Strauss, 1993). Grounded theory is premised on iterative coding of emergent themes, topics, and concepts. It is an empirical approach to qualitative data analysis that systematically constructs theory analyzing data. It involves moving between data collection and

analysis, constantly comparing themes, creating codes, cross-checking categories, making sense of puzzling findings by theorizing about them, and, throughout, keeping close contact with data and initial theorizing. The primary focus is on human activities, ultimately integrating categories codes into a theory within the substantive phenomenon. Here, the codes were analyzed for relationships, and themes were discussed in 3 rounds of axial coding (Elo & Kyngäs, 2008b).

In the first round, we read and coded the interviews with a focus on the work to deposit datasets, including practices of storing, collecting, documenting, analyzing, and sending datasets to repositories, resulting in initial codes and themes. In the second round, we reviewed the initial coded and added started looking into literature for explanatory theory and frameworks to explain two types of data practices found, selected articulation work coded for project levels of data practices. In the third round, we looked for empirical-theoretical connections to articulation work and coded for epistemic, material, and social aspects of data practices.

Where appropriate, the codes, themes, and memos were revised, e.g., renaming a code upon reassessing of the content, collapsing multiple code names into a single code, and/or splitting a code into additional codes to increase conceptual and semantic accuracy. From the coded transcripts, the categories were then generalized into supra-categories and subcategories. The supra-categories and sub-categories and themes found were related to the interview questions (e.g., *production work at the laboratory level*) but codes and themes also included emergent content (e.g., *data deposit facilitated by faculty in model organism repositories*). Throughout data analysis, the project team compared codes and discussed results.

5.6 Findings

We found faculty scientists are involved at many different project levels to publish experimental findings and get data deposited to federal and community research data

repositories. Data deposit occurs within goal-directed workflows, often in pursuit of a scientific problem that culminates in a published paper. Faculty scientists must engage in data work at multiple project levels and stages of the workflows to make sure journals, scientific protocols such as methods, and data repository criteria are satisfied. Because many genomics publishing venues require authors to share data, the workflow of faculty scientists to publish a paper is inextricable from preparing data for deposit (**Table 10**).

Table 10: Categories and Codes of the Data Analysis of Interview Transcripts.

Category/codes	Description
Work of Data Production	
Producing data from experiments	Generating data from techniques and technologies, often standardized, for laboratory experiments. This work helps to ensure data is accurate. (*Data reuse*, while not the focus of study, appeared in this category)
Organizing datasets for analysis	Datasets are managed through labeling, storing, documenting, systematically ordering data. This work ensures datasets can be shared.
Preparing datasets for publication	Reviewing experiment data and images.
Work of Data Sharing	
Identifying genomic repository	Searching for and learning how to use the genomic data repositories to deposit data. As changes occur, keeping up to date with the new standards and processes.
Communicating with curators	Exchanging information and engaging in dialogue with repository curators to algin the needs of the repository and data goals of the lab.
Depositing datasets	Ensuring data are documented based on repositories archival standards

Faculty scientists engage in data work from the level of data production up through sharing of data to a research data repository. Reusing data did occur during “data production” (**Table 10**), where faculty searched for data to compare with their produced sequences to identify biological function or significance. While workflows tended to sequentially unfold, we found faculty data practices are recursive, accounting for unexpected contingencies and to coordinate

with a research group of graduate students, technicians, inter-departmental and international collaborators, as well as data repository curators and staff scientists. We describe the work faculty do to get data deposited into a repository, from the work of data production to the work of data sharing.

5.6.1 Producing Data

Data production work processes in genomics commonly follow standardized methods, technologies, and techniques. A key objective of using standardized materials and methods is producing accurate data that retain their integrity across contexts. Faculty scientists in this study described data production work as following standardized sample preparation techniques and experimental analysis approaches such as protein immunoblots, x-ray crystallography, polymerase chain reaction (PCR), high-throughput gene expression, and single cell sequencing. Data production approaches often take place within a faculty's laboratory space and are performed by her research group. While it is not common for faculty to work 'at the bench,' faculty participants reported being involved in overseeing data production and working with research group members, e.g., graduate students, in the lab at the bench on data production. As P2, a studying gene expression related to disease development at a public research university describes, his work is to manage lab activities and supervise data work:

> *Some principal investigators manage to do a little dabbling in the lab. As a new investigator, you work in the lab most of your time, but as a more senior investigator, you end up spending time doing everything but bench work. So, it's the people in the lab that are generating the data, the post docs in particular and students and technicians. Their job is to generate data. So, they spend 90% of their time generating data.... (P2)*

In the lab, data is produced using a combination of high throughput data production techniques and more 'traditional' data collection methods. P11, a tenured evolutionary biologist at a private research university, described the prevalence and role of 'bucket and dipnet science' relative to the more and more common use of high throughput technology for generating data:

> *There's so much technology involved in doing most science. Not all. Like I said, we do field work on Dung flies. We've got, you know, plastic vials and a tube with a glass rod at the end that we suck flies into off of fresh cow dung. And we bring them in, we measure the length of their feet actually as an index of body size. We look at mating success, who's mating is not…. A lot of the science we do is still [...] 'bucket and dip net science.' Not everything we do costs a lot and not everything we do is very technically challenging.* (P11)

For P11, generating data in genetics and genomics is an amalgam of sophisticated new techniques, traditional methods, and approaches that lie somewhere in between 'bucket and dip net' fieldwork and emerging approaches, e.g., robotics and CRISPR gene editing. Outside of the lab data production, faculty manage contracted work sent to companies and sequencing facilities like the molecular analysis core which offer pay for a service data preparation and houses common-use equipment. Reflecting the multiple types of data production, P1, a tenured professor at a public research university studying epigenetic regulation of cancer protein, describes a typical workflow in sequence data production culminating in data deposit. Frist, P1 describes a traditional sequencing workflow:

> *We do [sequencing] at a couple of different levels. So, for most of the biophysical work we do we have genes that we've cloned and that we sequence. The genes are not terribly large, and we often send them out for sequencing, the traditional Sanger dideoxy sequencing methods. So those are sequencings that give reads of about 500 base pairs. That'll confirm the identity of our gene or if we've been successful in incorporating. A lot of times we'll change just a couple of base pairs, so that'll change the amino acid. Then we can study the protein that's made from this base pair change. That's the major workflow that we do for sequencing.* (P1)

Then, P1 describes a newer method for data production using next generation methods. Like the multi-level sequencing workflow P1 described above, the sequencing workflow for next-generation methods crosses multiple locations and traverses physical and digital mediums, from the faculty lab to the institutional core facility then back to the lab for manipulation using a software program to extract and process the data. P1's describes the data production as crossing through multiple stages inside and outside the lab:

> *We have started doing more big data kind of sequencing. I have a grant that's pending right now where we're proposing a combination of what they call RNASeq and ChIP Seq. RNASeq is where you would take cells — and we're working with human domain cells — you isolate all the RNA from them and then you want to convert it all into what we call cDNA, complementary DNA, that's based on the RNA sequence. And then we send it out for next generation sequencing. It comes back with recounts for all the genes that are present in that RNA sample. We'll add the inhibitors to sell and then look at how it changes gene expression, pattern of*

> *cells using this next generation sequencing kind of approach. So, we take the cells and often what we'll do is put them in a little cell pellet and we send them over to our what's called the [university] core facility, which is here. So, it's a facility, a core facility where they have people in there that'll take the cells, extract the RNA, convert it -- the cDNA — for us and then set up and carry out the next generation sequencing runs. And then they'll let us know when it's finished. And then we can log in and see 350 million pieces of data [laughs], which is a little overwhelming*. (P1)

Ensuring data is accurate during data production can be delayed and require work of correcting errors from contaminated data. For example, P11, a tenured professor at a public research university using genomics to study psychiatric disorders, described the revisions required to reproduce data:

> *I heard this story from a conference, saying for many years people cultured a HeLa cell. Then they come up with some kind of interesting finding and many can publish. They seemed to have some consensus, some common finding. But it turns out to be contamination. So, all the cell lines every lab was using were contaminated. Some person sees the problem and make a publication that says, "Every cell I used was contaminated." Then everybody's publication, they were all wasted. So, you can see this kind of small issue can be really devastating if you don't catch them early. Like in my example, I wasted thousands of dollars on that first experiments. We were completely confounded, we could not really cite that real data, basically we cannot use it* (P11).

The work involved detecting the anomaly and then alerting the wider community to the issue. The faculty had spent thousands of dollars on the experiment using the contaminated cell lines. Because the contaminated data was a widespread problem, uncapturable by the review because hundreds of papers were published and "the reviewer didn't know this or capture the problem" (P11).

While the data production method is specific to the research problem, producing data proceeds with standardized materials and methods. Across research contexts and problems, the materials, techniques, and technologies for data production are designed with systematic standards with a goal of producing reliable, accurate datasets. As P1 elaborated, the software *Base Space* is a tool scientists use for organizing and analyzing data:

> *Then we login to use something called* Base Space *which is, I think it's part of the Illumina platform. So, it's just kind of a nice place to keep your data organized. You can process it there. And then they have all these tools you can use for you know, extracting information and all that stuff.* (P1)

Whatever the approach, standardization is intended to ensure data accuracy and reliability. However, standardized approaches are part of a larger workflow wherein the tasks carried out are prone to error, both human and machine. Samples can be contaminated, techniques done incorrectly, or skewed by 'human bias.' The work contracted to sequencing facilities, as in P1 and P11 above, is removed from scientists' immediate scrutiny and control. Ensuring data accuracy involves working with contractors and requires faculty to vet the quality of both the process and products of the work. Data accuracy requires assessing the quality, integrity, and reliability of data produced inside and outside the laboratory. For example, P1 and P5, molecular biologists who work with the institutional core facility. P1 emphasized the importance of establishing trust in contracted data producers:

> *We're sort of new to the next generation sequencing stuff, so we've been kind of muddling along with it, which has been challenging for us. So, we will interact with the people over at the [university] core facility. One professor over there, very good, but he's been doing this for a lot of people. So, it's been tough to get a lot of the analysis done. We've been trying to learn the analysis on our own… [but] there's an energy of activation there because we're busy with so many other things. And I think that to do that really well, it has to be sort of your main focus. So, we'll probably collaborate with people who run the city [university] core facility. I've contacted people at other universities and formed collaborations with them. Once the data's done, I send it to them, and they start to do the analysis that way. I also recently used a service where we sent the data to a company that specializes in processing next generation sequencing data. The service was called Acura science and that worked out pretty well.* (P1)

By consulting with the core facility professors, collaborators, and third-party data producers, P1 checked the accuracy of the data when generated at a distance, whether as at the core sequencing facility or third-party company. The work of ensuring data is accurate in situations removed from production in the lab, faculty work with data producers and have conversations with them to glean information about the nature of the analysis. As P1 emphasizes, he has a sense of confidence in the accuracy of the data based on criteria he uses to assess the reliability of the data:

> *I feel more confident [sending the data to the core facility] because they've worked with a lot of people and had successful results. They also they don't mind listening to my stupid ideas and let me know when I'm right*

> *or wrong. I'll tell them, "Listen, I think that we should analyze it [this way], we usually approach the data with questions in mind, we'd like to use a hypothesis driven approach to doing this."* (P1)

Faculty scientists also work to make sure data is accurate by interacting regularly with the students who generate data. Where there is a particular vulnerable procedure, such as hypothesis testing analyses prone to bias, faculty create systems lab for doing validity checks. For one, faculty set up ways for data checks to occur by establishing expectations for the students to a) perform experiments correctly by setting up mentoring partnerships between undergraduate and more senior students who know the correct way to perform data production and analysis techniques, b) scheduling regular meetings and encouraging ad hoc meetings with the students and being in the lab with students to guide data work. As P8, an assistant professor using mouse mutants to study epigenetic and genetic factors linked to neuropsychiatric disorders, described her insistence that she be in the lab with students to "at least [be] very involved in looking at their data. She expressed how she prefers to be "physically in the lab doing experiments with [students]" to ensure data gets to the next level of analysis:

> *But I'd like to be in the lab at the bench beside them teaching them, doing experiments side by side and being involved in the actual hands-on. If not that, at least being very involved in looking at their data. Meeting with them very regularly. Having my door open so that they can come in when there's a question. And going into the lab and looking down the microscope with them. Looking at the results. Looking at the data. Working with them till we're through to the next steps.* (P8)

Faculty cannot always directly supervise data production in the lab. To manage the risk of data inaccuracy without their direct oversight, faculty create data accuracy intermediaries to serve as a check on the accuracy of data at multiple points in the workflow. For instance, P8 has her technician and mature student researchers act as a procedural checkpoint for the accuracy and integrity of the data production processes and resultant data:

> *So, if you have a technician or graduate student who has set up the experiment, they know what's what. So [they'll] help with the undergraduates who will have a coded set of samples. It's just that there is, humans will unintentionally bias themselves. So, we have samples where you have the control condition and you have the ones that were treated with vitamin D for example, you want that to work. You can't help it. And so, when*

> *you select which ones to measure, you might unintentionally bias yourself to skew the data. So it's best to have somebody who has no idea which one was treated and which one wasn't. Actually do the measurements and analyze the data so that they don't know which way. And then we break the code afterwards.* (P8)

To prevent biasing the data results, P8 designed data production and experimental workflows with data validity checkpoints by setting up technicians and senior student researchers to act as stand-ins. In lieu of the faculty directly assessing the process to ensure data accuracy and experimental validity, she invests her trust in the trained lab members to carry out experiments with the correct procedures.

In overseeing data work in the lab and the wider scientific community, faculty scientists also ensure data is accurate and validly produced by correcting students when students manipulate or falsify data. Faculty in our study reported their experiences with students who generated data and ran experiments then selectively extracted data to depict a positive correlation in the treatment condition. For example, P6, an associate professor at a private research university studying embryonic development using genomics techniques, explained the challenge of correcting manipulated data which her student had produced. P6 had instituted similar data accuracy check points in her lab as P8 (above), such as scheduling meetings to review student data and interacting with sequencing core facility scientists. However, the student in her lab felt pressure to publish positive results to be competitive with peers who may have amplified the significance of their data to publish in prestigious journals. In explaining the incident, P8 expressed her sense of responsibility to detect and ameliorate threats to the accuracy and validity of data produced and analyzed in her laboratory group. In the situation of the student who manipulated data, P8 spoke with explained how she addressed the threat to data validity:

> *I can tell you a disappointing conversation that I had with one of my students where we were writing revisions for a paper. The student runs the westerns, and then we quantify the westerns, and then I asked him so what do you find? [He said] "If I really run the numbers, it comes as statistically significant, but if I run it this other method, I'm still running statistics but I'm computing the data in a different way but then it is not statistically significant. So we can publish that one." And I'm like no, this is not the way that you do research.*

> *[...] Because he was just trying to use different programs to quantify the pixels and then quantify the pixels in different ways to just see if this will come out as not statistically significant. [...] We [had] this complete ethical discussion, where he was not operating in a malicious way. But he was feeling the urge to please me or, I don't know, to find something that he was not seeing. He was just so exhausted [by] the whole review process. He's like "I just want this published, and I just don't care anymore." At that point in time, I just asked "okay, you are totally biased on this so just give me the primary data and then I will look at what I see and then I will decide what is the best method for quantifying this" […].* (P8)

> *...So we decided to remove it completely from the paper. And then at that point, there were even tears in his eyes just saying that he knows that other people do it. And they publish very well. So if other people [do] it and publish very well, why not him? Why doesn't he deserve to have a Cell paper? We published it in a lower journal. But this was with tears in his eyes. I know that everybody does it. And if everybody does it and publishes well, why can't I? And I was terrified. (P8)*

P8 spoke to her student and to prevent further had the "difficult conversation" with him to address ethical concerns with data handling and manipulation. Other faculty expresses issues with detecting data accuracy issues and address concerns with data handling. For instance, P4, a tenured professor at a private research university using genomics to study plant pathology, regularly meets with students to discuss instances of questionable data handling or interpretation of experimental results:

> *It's easy to ... it's not even a falsification, right? It's manipulation of the data that leads you to a conclusion that is not the real conclusion. And that is very difficult to teach. I had this conversation with him. I had another conversation about ethics before where the same thing [happened]. I removed the whole section of a paper because he was showing me two movies that showed an effect. But when I saw the 20 movies that he had got and it's like look, if you see two movies that show this, that the 18 left don't show it, we cannot publish this as significant. It may well be that that's happening, but I don't have enough representation to say that that's what's happening.* (P4)

Faculty work related to data production involves ensuring data accuracy. Participants expressed how their work to ensure data accuracy spans a range of activities moving across spaces and different levels of work organization. The cases of P1 and P11 illustrated the ways data moves from the lab to external data processing facilities and back again, requiring the faculty to exercise vigilance by conducting validity checks and establishing a baseline of trust with external data processing entities, such as the core facility, collaborators, and sequencing companies. Another mechanism for ensuring data accuracy was creating checkpoints at specific vulnerable stages of the data production workflow and informal systems, e.g., meeting with

students, to ensure the accuracy of primary and secondary data produced. As demonstrated in P11, P8, and P4, creating a reliable data production process involves establishing checkpoints to act as formal and informal ways of ensuring data accuracy, integrity, and validity during data production.

5.6.2 Organizing Datasets for Analysis

Faculty scientists described the work of organizing datasets as crucial antecedents to making data *deposit-able*. When datasets were well-organized, it was easier for research group members to access and share datasets. The results of data analysis retained greater interpretability if the datasets were organized through standardized documentation and archiving methods. Unlike data production, methods for organizing data do not reflect a standardized approach to the same extent as, e.g., the western blot or Sanger sequencing methods. To address the relative lack of uniformity in procedures for organizing datasets, the faculty scientists described the ways in which they set up data organization in their research group. Access to data is crucial; and faculty do work organizing data to ensure access is possible, within the lab group and later for distribution outside the lab and depositing to repositories. Faculty organize their research data to prevent degradation of samples, materials, and datasets. Faculty set rules for data organization work to ensure there is at least one physical copy of the data that remains in the lab. As P2 describes, she instructs her research group members to only take a copy of the data but organizes data to remain ‘within the lab and with the grant’:

> *The data stays within the lab, with the grant. I let people take a copy of anything they want. A lot of times they need especially their protocols. That's what they really want. But the data needs to stay here. A protocol is a method for how you're doing a certain type of experiment, isolating the subcellular fractionation. I'm like take all the protocols that you want and you can copy some of your data. Sometimes they're still writing papers when they leave. I'm like, just make copies but original stays here.* (P2)

P2 also instituted a system for data and experiment documentation in her lab through the

regimented use of a lab notebook. The laboratory notebook is marbled or spiral bound book that is stored on shelves in the laboratory. Because the data is derived from samples, organizing the reagents and other materials is important to create a provenance trail of not only data but the original samples and their metadata using the notebook.

In addition to the work of organizing datasets, documentation, and samples, faculty develop file systems to manage the regulatory and policy dimension of data. For example, P9 entered into Nondisclosure Agreements (NDA), Material Transfer Agreements (MTA), and a Memorandum of Agreement (MOU) with biotechnology firms that detail the limitations for data use and indicate how data should be securely stored. As she described, the legal stipulations impact her work of organizing the datasets because they had to be carefully stored and not shared with outside entities. P9 describes the dynamics of organizing the data within the institutional and legal stipulations:

> *If it's with a company, then you're going to go through huge non-disclosure. That's what I went through, I realized the lawyers who had to be involved before they'd even have a conversation with me. So that was very clear I couldn't even talk to them until the lawyers went through it and made back and forth a few times. With most collaborators, you discuss the parameters going forward or... often it's quite well understood though that if I give you data, you don't pass it on to somebody else. And usually common courtesy would say, "I'm presenting this seminar at this conference, do you mind if I talk about this project?" So it's usually not formalized. I think if it's something that's involved in technology transfer or drug companies, formal agreements are usually in place. If it's general basic science, it's usually just an agreement or just a general understanding that if it's their data, you don't present it unless you have their permission informally. It's a professional understanding.* (P9*Now if it's a reagent, for example, usually the universities insist on a material transfer agreement that outlines these things, an MTA. For example, if somebody made a transgenic mouse in my field, a mouse knockout. Now often these days people make it freely available, but some don't. So if they've made a knockout mouse and they're willing to share it with me, I will likely sign an MTA and the universities will have agreed. It has to go through their office, the IP, their actual property office, who will usually say, "Okay here are the rules, so it says basically that I have to give them credit on the publication or I can't send it on to somebody else without their permission."* (P9).

Organizing datasets for analysis also requires knowledge of how to label and describe datasets. To do this, faculty develop protocols and enforce standardized procedures for organizing data, such that analyses can be rerun if necessary and interpretable after the data is analyzed. The organization of data includes keeping digital and physical copies of data and data

documentation in a manageable format. P3 describes her work process of creating both a paper version and a digital copy for longer term storage.

> *We make a strain, we have a paper version, cause it's got a little more detail and people are supposed to make out the paper version and then it gets loaded into the Excel spreadsheet. We have one for yeast, one for E. Coli, which is how we manipulate DNA. And one for oligonucleotides that we've ordered. I can't say it's perfect, but it's actually pretty good. We have thousands of strains. And then there's a, a number that's assigned to each one that tells where to find it in the freezer. We can freeze those strains. So we take a little bit out and revive it and then send it.* (P3)

Each medium has its advantages and disadvantages; the paper copy contains additional information whereas the digital version in Microsoft Excel has a better chance of longevity and discoverability in their database. Part of data organizing work includes managing compatibility with institutional infrastructure. For example, the labels for the sample provided by a company, or the servers and computing software provided the university.

For example, P3 uses the labels from the sample company to describe their data. The labels are a shared standard that enables data to be organized for use inside and outside the lab. For instance, the information from the company is a naming convention that would be recognized by her lab group members and the company if the lab needed to contact them to troubleshoot an issue. As P3 describes:

> *And so we just put in the information from the company and try to name it in a way that would let people search and actually group them by the gene it was targeting.* (P3)

Instituting the naming convention for the data is part of work of developing a series of interconnected steps to make sure data are accessible and shareable within the lab, enforcing data management procedures in her group, and adjusting to university systems, such as with the use of login credentials. Similarly, P6's data organizing work involves continuous updating and enforcement work so data can be made accessible and shareable. As P6 describes, she tries to enforce a standard series of steps protocols are sharable:

> *I told everyone you have a new protocol, you scan it, you put it there, and it's there for everybody to know how the lab does this thing. I set it up when I came here and I forced everyone to use it, and there's these, I don't know, spring cleaning episodes that you get. Like, "Let's organize the server." Right? And then I do*

> *it, and I push everyone, it's like, "Hey, I haven't seen that protocol. And I haven't seen that protocol, and I haven't seen that protocol, so you upload them please, before the end of the week." Then, with that surveillance, it happens, but I find it that it's so much more easy to just ask people, "Take a photocopy, and put it in the binder." They would rather do that than do the server. Don't ask me why. It does take the same amount of effort, because the scanner is also our printer, but physical things seem to be easier to maintain for people, somehow.* (P6)

Another goal of data organizing work is to reduce redundancy, e.g., buying materials twice. For example, P3 described her data management and sample organization systems as 'not perfect,' but pragmatic to prevent purchasing an excess of materials:

> *It means everybody has to make out the sheet, right. It has to get put in there, but that's one of the things the lab manager works at is keeping that up and for oligonucleotides targets. So those we buy and it's mainly just making sure that we don't buy something we already bought.* (P3)

Faculty like P3 and P6 reinforce the procedures through communication with their lab group via weekly meetings and lab handbooks that explain the expectations for data organization. They also rely on technician, lab managers, and the apprenticeship-style model to teach junior lab members. In other words, faculty organize the lab such that the varying levels of researcher seniority, e.g., a mix of undergraduate, graduate, and postdoctoral researchers, enables communication between students to teach and reinforce norms of data organization.

The purpose of organizing data in the lab is not only to satisfy immediate data requirements (e.g., for sharing or facilitating data interpretation), but also constitute an effort to prevent data loss. P3 intimated there is a debate on the use of electronic lab notebooks because of worries about data loss.

> *The [scientific] field is just: "An electronic lab notebook?" I just am worried about… I think people haven't thought through that. You know, I can't even read a [Microsoft] Word file from 10 years ago. And these lab notebook programs are quite specific, if that company goes out of business are you going to be able to read it?* (P3)

Another data loss contingency can occur if a student leaves the lab with the data. As P13, a tenured professor at a private research university studying fish genomics, expressed:

> *If I give you the most important project, funded by our government, two years later you say I'm going to leave your lab. I'm screwed. Not only this project will be delayed, but also if you generate a lot of data and I say, where is the data, you say I'm not going to give it to you. Worse scenario, what can I do? I can feel that you are stealer. That's about it, what else can you do. It's better science, even safer to have put more, at least two students on to the same project. One may be leading and one may be engaging and almost equally leading but secondary leading so let's say you have all the data. Another student, when she is leaving, I say, at least you transfer all the data to the lab. You say no, I don't care, but at least I know, as a back up position. So you understand, at least that eases a lot of tensions.* (P13)

Faculty also organize data in anticipation of future access and interpretability needs. For example, P6 developed the protocol storage system on a server with longer-term data organizing goals in mind, namely, to control and manage access to the protocols for students in her lab, even as students came and left the group. P3 similarly created a standardized procedure for keeping samples organized, enforcing its use, and instructing her lab technician to maintain and check the database. In sum, faculty create systems within their lab for continuous organizing and documenting data in anticipation of needing to access the datasets later, to prepare a publication, e.g., or deposit data in a repository.

5.6.3 Preparing Datasets to be Publicly Released

P2 is a professor of biochemistry and molecular biology at a public research university whose research focuses on cellular stress response with yeast as a model organism. P2 instituted procedures and enforced more informal expectations within her lab for students and researchers to create a provenance trail of data to demonstrate the accuracy and reliability of data they produced. For each experiment, P2 requires students to create a spreadsheet of timestamps to show the number of experiments done and when the samples were produced. ‘Getting the data right’ was how P2 expressed her efforts and commitment to ensuring data accuracy. Her

approach to assessing data accuracy is through visually inspecting data and documentation, including lab notebooks, that passes her assessment of 'getting the data right.' In P2's lab, she discusses expectations for data to be ready for a publication.

> *Within my lab, we talk about [when the data is ready for deposit], but also I think when you do your first papers, I want to see the data and things like that...how many times did you do that? I like an Excel sheet that shows what the dates were of the time she did it.... How did you do those statistics?* (P2)

By showing P2 the provenance chain of data production, evidenced through the western blot images, timestamps, and subcellular fractionation results, the validity of the data is established. Data satisfy the threshold for moving closer to the repository. The lab members, generally students, curate the data provenance chain of evidence by pointing to standardized method, e.g., the western blot, in a lab notebook.

By reviewing experiment data and images, P2 and other faculty members establish data are ready for publication and deposit in a repository. The data is ascertained to be in line with the publications claims and "pointers" to the data are verified so that if reviewers request data, the lab members can provide it. Similarly, P3, a professor at a public research university studying gene expression, remarked on the work of generating more data at the journal reviewer's request. Preparing datasets for publication involves working in the lab with hands-on experimentation but also negotiating and communicating outside of the lab, e.g., with the journal reviewers and data curator. As P3 explains, it took almost a year to meet the data requests of the reviewer:

> *So in fact in this paper we ended up including about half the data that we're going to use for the next paper. So in order to get it accepted, I had a different person working on the PR version, so the main author had done the body of this work and then there was a separate person doing a separate body of work and then we had to put this into there in order to get an accepted. So we ended up taking half this person's data project and putting it into this paper to get it accepted. When he went to publish the next paper, half of the data was already in here. So we had to take it even further to get that one published. So it's just like this cascading effect. So often they want more, there's a high, the reviewers typically want more data.* (P3)

Preparing data for publishing is work faculty do across research communities to satisfy the demands and requirements from multiple parties. For instance, the lab group of P3 was

tasked with extra experimental work and producing more data to address the reviewer concerns. P2 instituted systems to locate data and provide it for the data curators through Excel spreadsheets and documentation in lab notebooks.

5.6.4 Sharing Data

The work of data sharing involves laboratory coordination tasks, revisiting experimental work, and communicating with database curators. Faculty deposit data to a variety of genomic repositories (**Table 11**).

Table 11: Genomics databases where the participants deposit data

Data Repository	Participant	Description	Institution / Institutional Requirements
GenBank	P1-P11, P15	Hosted by the National Center for Biotechnology Information (NCBI) and an aggregator of data from other repositories (e.g., SGD, FlyBase, Mouse Genome).	Federal repository with professional curators stewarding data and data services. Major journals and funding agencies require data be deposited to PDB. Data must conform to data and metadata standards.
Saccharomyces Genome Database (SGD)	P2, P3	Model organism database for budding yeast.	Model organism database. Non-federal database staffed by programmers and biocurator scientists. Requires data conform to metadata standards.
FlyBase	P1, P3, P10, P7	Model organism database for *drosophila melanogaster*, i.e., fruit flies.	Model organism database. Community run / volunteers/ metadata standards
Mouse Genome Database	P8, P9	Model organism database for *mus musculus*	Model organism database. Community run / volunteers/ metadata standards
bioRrxiv	P1, P4, P18	A preprint server for biology, (following the model in physics of *arxiv*) to post unpublished papers and datasets.	Owned by Cold Spring Harbor Laboratory. No peer-review process, but basic anti-plagiarism screening checks.
Gene Expression Omnibus (GEO)	P9, P11, P12	Gene expression profiling database. HTP screening genomics data from microarray or RNA-seq experimental data.	Hosted by NCBI. Must conform to *MIAME* format standards.
Protein DataBank (PDB)	P2, P7, P10	3D structural data of biological molecules, e.g. proteins.	Managed by the Worldwide Protein Databank (wwPDB). Major journals and funding agencies require data be deposited to PDB.

Faculty deposit data to genomics as part of their workflow at many different stages in the publication process. Data repositories fall into multiple categories, each requiring deposited data to conform to the format and other criteria to make data compatible with others' submitted data.

For example, faculty navigate data deposit by learning how to use policies that are different according to the database and journal. As P11, a tenured professor at a public research university using genomics to study neurodegenerative diseases discussed the use of an embargo to release data according to a timeline decided by the database and the journals:

> *I think that it's often dictated by the journals themselves that publish it. They'll have requirements for when you have to submit those data to databases. So they won't release your publication until you've submitted the data. And then you get an ID number. You have to show the journals the ID number. [...] I think there's mechanisms by which you can submit the data and create an ID for it, a unique identifier, but that data will not be visible to the public until you get the okay. To some, I think there's some mechanisms by which you have to submit the data. It has a release date, you work it out with the journal. So as soon as the paper comes out, the data is released.* (P11)

Databases differ in their policies, but also how well-resourced they are and the extent to which they control data deposit. That is, databases with more resources such as NCBI's GenBank or the SGD at Stanford have bio-curation and programming staff scientists who are employed to administer and manage the data. The work of data sharing involves keeping up with the latest news and knowledge of administrative and functional components of the databases. As P2 describes, databases are essential to his work in genomics. As such, P2 demonstrates his work to keep up with how the model organism databases are managed and funded impacts the quality and content of services, resources, and products available for genomic research:

> *There's two really important databases that we basically cannot live without. And these are having trouble getting funded if you can believe it. So first we worked with yeast and Drosophila. The first database is for yeast and it's called Saccharomyces Genome Database, SGD at Stanford. And this is something that we basically cannot do our research without this site, SGD. There's a few people like at Stanford and other places that developed that, like the pioneering sites, like the yeast genome database to be very sophisticated and they talk a lot to each other. They try to make these databases as good as, say, the yeast one.* (P2)

Likewise, faculty learn to use the database features and policies, for instance, to protect their data, e.g., through requesting a 6-month embargo to delay the release of their data before the publication is accepted and finished. As P4, a tenured professor of plant pathology at a

private research university, explained, the work of sharing data with the repository requires multiple interactions to indicate the researchers' preferences, e.g., with respect to data release:

> *As soon as we generate [data], right before we write the manuscript and submit it, nowadays we are more conservative, so we put a timing limit. Say make it publicly accessible after six months. Or publication of the data, whichever is first. As soon as your data is published, it becomes public, or after six months you feel that it's very secure, but on one or two occasions we had to extend that another six months. Before they release to the public, they ask you, is it okay? Now you say okay. If you do not respond, their default is that they release it.* (P4)

The work of submitting data is recursive curation work that involves preparing experimental data, coordinating, and planning within the lab, and interacting at many different points to deposit data. P11, P2, and P4 indicated, faculty work to learn about different repositories. The work of preparing datasets to share is a data curation task but also requires updating one's knowledge about the database requirements. The work of sharing data to a database is ongoing, rather than a single act of submission. The same dataset can undergo many different formatting and packaging tasks to satisfy the different repository criteria.

5.6.5 Depositing Datasets

Depositing datasets requires faculty to perform data documentation based on repositories archival standards. This involves communication with curators and shuttling between work in the lab and interactions with repository staff. Part of the work of depositing datasets is becoming familiar with database requirements, e.g., changing metadata standards and reading documentation. These requirements become increasingly integrated into the lab workflows to the point that they are streamlined and 'second-nature.' As P4 described the work of submitting to GenBank, the repository has a standardized form for data and metadata entry, which enables heterogeneous data to be aggregated:

> *It basically has a standard format. Which organism, how did he isolate the DNA? How did you prepare it? What was the age of the plant? And things like that. And you just fill those things out, but that's not that*

> *hard. It used to be earlier when there was not a universally accepted format. Now there's a universally accepted format and everybody uses that format. So people can have comparable results.* (P4)

P4's case illustrated how the work of depositing data often involves work to learn and meet repository and institutional standards. Some of the requirements are ostensive and predictable, such as what metadata to include with the deposited dataset, such as how to classify the data. P4 described his data, a plant model organism called A. thaliana (the thale cress), it is rare for there to be ambiguity about the data such that the data submitter would struggle to describe the deposited data:

> *I classify this plant as under heat stress and somebody else may say, no it's not. So it depends on who is submitting the data. There are fairly straightforward questions, which you have to be an idiot to mess around with it. But it will all depend on who is submitting the data. But the questions are so straightforward that it should not be a problem you should be able to just easily be able to [submit]. Again, this depends on what type of data.* (P4)

In contrast, depositing data might require faculty to perform unanticipated data work. For example, P2, a tenured professor at a public research university studying gene expression in development and disease, described the additional experimental work required to deposit the data in a repository. The curators wanted to see negative data, in addition to the significant results:

> *I wouldn't submit until… Usually what they want is, 'cause you've published something, they would like to see the negative data too. And so that didn't necessarily go in that they want to see what you've screened and what was negative and so then I have no problem with. I think they just took like an Excel spreadsheet of, so the kind of thing is like the, so there's a big database where they've individual genes, you can buy this array of yeast strains with individual genes knocked out. And so they [the Saccharomyces Genome Database] will pull it from your paper when you, if you submit. But other times they want more data and I've just sent them, I think I just sent them Excel sheets. Yeah. And then they kind of mine it and link it to different genes and things like, so they have a page for every gene. It's very, very nice. And then phenotypes and things like that. So I think they pretty much feed your data into their format, but they just want data.* (P2)

Databases such as the preprint server bioaRxiv create options for where faculty can deposit their data. However, institutional requirements constrain their options to the extent that institutions currently endorse mainstream repositories such as GenBank, PDB, and dominant model organism repositories (e.g., FlyBase, SGD). Faculty are incentivized, then, to deposit data

in the prominent repositories and to conform to the formats, standards, and other requirements set by the repository.

5.7 Discussion

The findings show faculty engage in articulation work, that is, the work of aligning levels of work organization and work types to make data deposition to a repository doable. We argue that the data work of genomics faculty scientists to deposit datasets is work to align the epistemic, material, and ethical aspects of their data. This work is articulating alignment in the sense which Joan Fujimura (1987) outlined in her model of how scientists construct a do-able problem in cancer research.

5.7.1 What We Mean by "Articulating Alignment"

We adopt Joan Fujimura's (1987) alignment concepts to uncover the articulation work of data deposit. Fujimura conceptualized the do-ability of scientific problems as the alignment of three levels of work organization (experimental, laboratory, and social world) and two types of work (production and articulation) (Fujimura, 1987). These concepts help us to see the material, epistemic, and ethical components of data practices which faculty scientists in genomics engage in to get their data into a repository. We show that implicit to the work of making data depositable is the recursive interaction between the social world level of the data repositories and the laboratory and experimental level of the scientists' work.

In her model of articulation, Fujimura defines *do-ability* as the perception of scientists as to if a scientific problem is 'intellectually interesting" (Fujimura, 1987: p. 257). Scientists in Fujimura's study attributed doability to sophisticated new technology, claiming that emerging recombinant DNA techniques led to a surge in oncogene research because the 'productive methodology developed' enabled scientists to ask and address new questions, e.g., 'are there changes in cellular proto-oncogenes in tumor cells?' (Fujimura, 1987: p. 258). However,

Fujimura argued that technology alone is not sufficient to make a scientific problem do-able. Rather, Fujimura developed a model of doability that conceptualized feasibility as aligning work tasks through organizing and reorganizing work across project levels: 'Doability is better conceptualized as the alignment of levels of work organization' (Fujimura, 1987: p. 258).

In this conceptualization, there are three levels of work organization scientists need to bring into alignment to make scientific problems feasible. These three project levels of work organization are the (1) experiment level in which a set of tasks are performed in the laboratory, the (2) laboratory level as a collection of several experiments and other tasks like purchasing laboratory equipment (e.g., an ultracentrifuge) and the (3) social world level , that is, the broader social milieu in which experiments and laboratories are situated (Gerson, 1983; Strauss, 1978). Scientists accomplish alignment by articulating tasks across project levels.

Articulating means 'considering, collecting, coordinating, and integrating' between the levels of work organization Fujimura, 1987: p. 258). In other words, scientists make problems doable through the ostensibly quotidian daily practices of 'organizing and reorganizing their work' [italics in original] (Fujimura, 1987). Fujimura's (1987) model of doability emphasizes that advanced new techniques can misdirect our attention away from these mundane, taken-for-granted tasks of scientific work, such as "washing pipettes and signing up to use the ultracentrifuge" that enable projects to work out (Fujimura, 1987). Further, if scientists 'package the articulation work between levels,' aligning levels is easier (Fujimura, 1987: p. 258).

5.7.1.1 Three Levels of Work Organization

Classifying the tasks involved in scientific 'problem-solving' into three levels of work organization and two types of work is a useful way to conceptualize doability. The first level is the experiment, that is, the collection of activities or tasks performed in the laboratory for a given project. An example of a task at the experiment level might be running a western blot, also

known as a protein immunoblot. The western blot is widely used as an analytic procedure for detecting proteins in a tissue sample or other material by inducing the sample to undergo denaturation. The second level of work organization is the laboratory level in which many different experiments and other activities are performed (Fujimura, 1987: p. 258). For example, the laboratory level might involve scientists pulling together a spreadsheet of statistical results, meeting notes, and experimental images into a set of slides to present at a conference. The third level is the broader social world situating experiments and laboratories. The social world of molecular biology and biochemistry is the larger field in which experiments on tissue samples of a model organism are situated. Researchers, colleagues, sponsors, institutions, and other actor with similar concerns or who are all occupied with a shared family of problems constitute the social world.[7]

5.7.1.2 Two Kinds of Work: Production and Articulation

The two kinds of work involved in Fujimura's (1987) model of constructing a doable problem are distinguished by the nature of the task and, in part, where in the level of work organization they occur. Production tasks occur within every project level. What makes a task production work is the relatively well-defined, procedural nature of the activity. It is a standardized set of procedures and materials. Examples of experiment level production tasks in a genomic research laboratory is when a technician runs a western blot. Production tasks at the laboratory level might include activities such as buying reagents or equipment, choosing what experiments to perform, and writing grants. Like the experimental level, the laboratory level production activities are relatively formulaic. Buying new equipment, for instance, is a

[7] The concept of the *social world* is one Fujimura (1987) explicitly draws from Strauss (1978).

standardized series of steps where the materials and procedures are known and the steps well-trod and straightforward. At the social world level, production tasks include organizing a workshop for an annual conference and meeting with a funding agency program officer to discuss a proposal.

Articulation is the work of pulling together the production tasks to accomplish the project goals. Per Fujimura's (1987) model, articulation tasks are carried out between levels of work organization. Articulation tasks can be recognized as the planning, organizing, monitoring, evaluating, adjusting, coordinating and integrating activities that bring together production tasks, and consists of both 'planning and coordination...of production tasks long before they are to be done, as well as ad hoc decisions made at the time the tasks are implemented' (Fujimura, 1987: p. 260). As such, articulation activities involve what Schmidt later called 'first order and second order articulation' (Schmidt, 2002: p. 27). First order articulation work is considered the planning and coordination of production tasks in the stages before researchers perform them, the ensemble of 'independent actors constituted by a system of interdependent activities' (Schmidt, 2002: p. 27). Second order articulation work refers, broadly, to the ad hoc actions taken by researchers in the course of performing the tasks, often in confronting unexpected events or contingencies (Schmidt, 2002: p. 27).

The articulation in Fujimura's (1987) model connotes a sense closer to the second order, making reference to work by Strauss' conceptualization of articulation work as 'integrative organizational processes' (Fujimura, 1987: p. 260; Strauss, 1988) . In his work, Strauss (1988) makes an important distinction between articulation work and articulation processes and extends

negotiated order approaches[8] to organizations, a conceptualization Fujimura (1987) inherits. As such, the distinction between two types of work and three levels is a pragmatic rather than categorical decision, evident when Fujimura suggests the importance of context in making an activity a ‘production’ or ‘articulation’ task (Fujimura, 1987: p. 260). In her words, ‘one person’s emergency is another person’s routine’ (Fujimura, 1987: p. 260).

5.7.1.3 Doability as the Alignment of Levels of Work Organization

Fujimura (1987) argues ‘doability is the alignment of the three levels of work organization: experiment, laboratory, and social world’ (Fujimura, 1987: p. 261). To align the levels, scientists ‘tinker’ within and between them (Fujimura, 1987: p. 261). Tinkering is another term for articulating. Fujimura (1987) employs the term ‘for its visual and hands-on connotation, in order to emphasize the dynamic construction of scientific problems within a particular context’ (Fujimura, 1987: p. 261).

The tasks at one level (e.g., at the experiment level such as running a machine learning algorithm) is only a subset of the total set of steps required for work to work (Fujimura, 1987: p. 262; Suchman, 1996) . Scientific work is accomplished when all the necessary components at all levels of the project are integrated adequately, ‘collected and made to fit together’ (Fujimura, 1987: p. 262). Scientists not only decide which parts of the workflow are necessary, but then also gather them to ‘craft and carry out’ a working scientific problem. According to Fujimura, alignment constitutes the work of making a problem doable. That is, by determining what tasks are required and then assembling the set of tasks — by shuttling between levels of work organization and reducing friction between levels — is crafting a scientific problem (Fujimura,

[8] As such, we interpret Fujimura's (1987) model of articulation here as avoiding the assumption of a tight, machinelike integration of work but rather an extension of Strauss’ negotiated order approach to organizations (Strauss, 1985, 1988).

1987: p. 262). “Problems are more or less doable depending on how difficult it is to articulate among levels to create alignment” (Fujimura, 1987: p. 262).

Technology alone is not enough to ensure a problem is doable or other scientific goals, e.g., getting published, are accomplished. Fujimura’s model brings to the fore the tasks of gathering and coordinating, the articulation tasks, to show how they make projects ‘work out’ (Fujimura, 1987: p. 262). Fujimura’s study participant, the Molecular Biologist who ascribed technology enabling a ‘productive methodology’ with the responsibility of making a problem feasible, focused on production tasks (Fujimura, 1987: p. 258). Like infrastructure which tends to be invisible until breakdowns occur (Star & Ruhleder, 1996; Star & Strauss, 1999), we notice articulation work when processes break down. Here, Fujimura focuses on labor that makes scientific work *work*, rather than when ‘things don’t work out’ (Fujimura, 1987: p. 262).

Alignment makes things ‘work out’ (Fujimura, 1987: p. 262). Traversing all three levels of work organization is necessary for a scientific problem to gain traction and reach a point of success. Articulating between the experimental level and the laboratory level is not enough; the institutional demands (social world) must be met for do-ability to obtain. Fujumura (1987: p. 262) uses the metaphor of a stack of paper transparencies to illustrate doability as alignment. At each level, work tasks are organized and fit together according to the logic of work practices of an individual, the laboratory, and the division of labor. If the transparencies can be rotated so that the current configuration of tasks at each level algin, then a problem is doable. Another way to align the transparencies is to tinker with a task within the level, that is, adjust or reconfigure the work to align with criteria, demands, constraints, or other contingencies at the other levels. For example, adjusting within a level can include discontinuing a project, switching out one task for

another, changing a task's structure, 'relabeling the problem,' or 'substituting audiences' (Fujimura, 1987: p. 262).

5.7.2 Data Deposit as *Doability* in Genomics

Data deposit in genomics is made doable by the work of articulating across and between three levels of work organization and two types of work. Production tasks in the workflows of genomics faculty include entering dataset metadata when depositing, performing experiments (e.g., western blot, x-ray crystallography), and sending sequencing jobs to the core facility. Articulation tasks include, for instance, correcting data contamination, learning repository guidelines, and making sure data is not redundant with existing data. Faculty explained how their data depositing workflows occur at all three levels of work organization described by Fujimura (1987), each of which involves production tasks and articulation tasks.

5.7.2.1 Types of work and projects levels

At the *experimental level*, the faculty participants in the study carried out tasks to produce data, such as acquiring and preparing samples, sequencing data either in-house or contracting the work to a core facility or company, processing the data, and extracting the information needed for further analysis. The production tasks included the standardized materials and methods, such as clones, the western blot technique, and PCR. Articulation tasks at the experimental stages of data generation included the within-level work of detecting and correcting data anomalies (e.g., P11's contaminated sample) and the intra-level work of communicating with the core facility (e.g., P1's questions to staff scientists about uncertainties with experiment design). To encourage data accuracy in the data production process, faculty set up data quality checkpoints, often in the form of a senior student or lab technician, as an authoritative proxy for faculty who would

usually directly check data quality, e.g., experiment data produced by junior lab members. **This work ensures datasets are accurate**.

At the *laboratory level*, faculty carried out production and articulation tasks to organize data for analysis. The work of organizing data included specifying procedures for data documentation, the storage of experimental protocols, and labelling samples and datasets, and managing access to data. Production tasks included the standardized procedures of data management, such shared protocols for entering information into a lab notebook and who is allowed to copy experimental protocols. To link production tasks together at multiple levels, faculty performed articulation work to align the tasks necessary to organize data. To organize data to ensure it was finable and interpretable within the lab, faculty ensured experiments were documented in lab notebooks (experiment level) and the experimental labelled based on company information provided about the sample (social world), while enforcing the proper storage of labelled data in freezers (laboratory level).

The intra-level alignment work of organizing data required faculty to align the three levels — experiment, laboratory, and social world — and developed ways to encourage and enforce alignment, such as by creating a lab handbook that detailed the data management procedures within the lab. Sometimes, tasks had to be tweaked within the laboratory level to align data organization tasks there with other levels. For example, when P2 had a publication reviewer request more experimental data, the scientists in his lab group needed to tweak the experimental level of data organization — generating more data, documenting the data, adding it to the database (at the experiment level)— then incorporating it into their paper (at the laboratory level) to be reviewed by the journal (social world level). Here, preparing datasets for publications crosses all levels. Organizing datasets for analysis ensures datasets can be shared within the lab.

Preparing datasets for publications involves reviewing experiment data and images to verify and check data completeness and quality. **This work makes data interpretable and accessible to share within the lab**.

At the *social world level*, faculty scientists engaged in intra-level alignment work to share data to repositories. The work involved interacting with different kinds of genomic repositories. Faculty communicated with biocurators and staff scientists to align the data production work with the repository requirements, learning and adhering to repository standards and expectations such that data met thresholds of repository standards. Production tasks at the social world level included entering metadata into the repository form. **This work ensures data are documented based on repository archival standards.**

Each process area constituting a data deposit workflow can be conceptualized as primarily falling into one project level. For example, the process area of *producing data* primarily falls under the *experiment level* of work organization. The typical tasks done in the process area of *organizing data* can be classified as falling into the laboratory levels, and *depositing data* generally is considered at the level of the social world because it is within the wider field. However, all types of work — from producing data to depositing data — implicate all project levels, e.g., experimental work involves using standardized instruments by mobilizing knowledge constructed and maintained amidst a broader social world. The purpose of foregrounding the articulation work of alignment project levels aims to draw attention to this, especially its recursive elements.

5.7.2.2 Work across project levels

Through the workflows of depositing data, data work in genomics requires continuous and recursive work to be made deposit-able. Fujimura (1987) describes the ways in which the levels of a project are made to fit together by a) deciding what are the necessary tasks and b)

adjusting within and between levels to gather tasks together in a set. Likewise, the faculty in our study demonstrated working across project levels, engaging in both production and articulation work to make data deposit '*doable'* through the alignment of project levels (**Figure 9**).

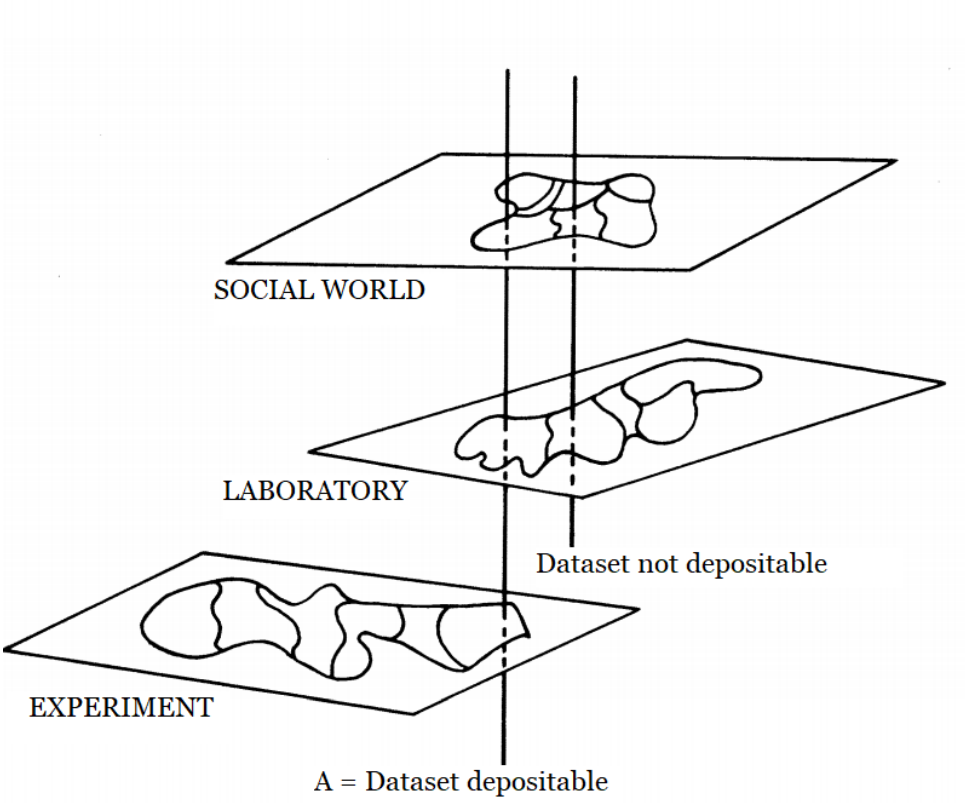

Figure 9: A metaphor for aligning data deposit tasks at three levels of work organization: experiment, laboratory, and social world. Adapted from Fujimura (1987: p. 263).

Scientists assess data at every step of the workflow - Faculty data practices traversed project levels and included passing judgement on data at multiple steps in the workflow. Where they could not be directly involved to assess the quality of the data, faculty instituted proxies to verify junior lab members produced, analyzed, organized, and deposited data correctly (P1, P4, P6, P9, P13, P14, P17).

Verifying work - To 'get the data right,' in the words of P2, who insisted on documentation of datasets and a controlled vocabulary for labelling samples (see Findings section). Faculty

monitored work at the experiment level by engaging with students at the bench and then holding regular and ad hoc meetings to 'look at the data' directly at the laboratory level (e.g., P1, P6, P8, P9, P10, P12, P15).

Making sense of data for experiment work - Experiment work traverses from the lab where clones are made to the sequencing companies or core facility then back to the lab. Making sense of the data requires work across levels. At the laboratory level, faculty design experiments with lab members based on problems defined at the social world level of the field. At the experimental level, faculty work with students to view the data and check the accuracy and completeness of data received from the sequencing service. When the data is unclear, faculty call the biocurator of the source of the dataset or the core facility staff scientists. They gather the tasks at multiple levels and fit them together to make sense of the data, as evident in P1's interactions with, and relationships with, core facility scientists. The distributed nature of sequencing data production makes this kind of work across levels — and the requisite alignment work — a frequent type of work in genomics (e.g., P1, P2, P3, P4, P5, P8, P9, P12, P18).

Judging when data is acceptable to publish - Preparing data to produce a publication involves work across project levels. That is, faculty judge when data is acceptable for publication by shuttling between experimental, laboratory, and social world levels. For example, P4 explained a situation in which the origin of data came into question. To make data depositable and publish the paper, P4 described holding back from submitting data until her lab and her had generated negative data and additional experiments to satisfy the requests of the data curators (e.g., P2, P4, P5, P10, P17).

Work across project levels illustrates the work underlying research data quality, completeness, interpretability, and integrity. What evades popular and policy narratives to

support and enable AI in genomics is the backend data work of scientists who create datasets that feed into the cyberinfrastructure and subsequently machine learning models. Understanding and creating a journey for pooled datasets through cyberinfrastructure as a marker for its value for AI genomics misses the point that datasets have interpretative flexibility which begin with the creator of the dataset (Leonelli, 2014).

5.7.3 Enabling Data Deposit: Aligning Epistemic, Material, and Ethical Dimensions of Data

Across the many different workflow steps of depositing data, it was necessary to shift between levels of work, e.g., to correct data issues. To deposit data, we found scientists organized and re-organized their work to meet the requirements and various demands at each level of work organization (i.e., the social level, the laboratory level, and the experiment level). We found constituent of data articulation work are several considerations, or what we call *thresholds*, for data deposit. These "thresholds" are enacted to make data deposit do-able. We found there are three types of considerations that must be satisfied for data deposit to be made do-able: epistemic, material, and ethical. Building on the idea of data articulations and these thresholds, we coin the term *datarticulations* to describe how scientists made data depositable through dealing with. *Datarticulations enable scientists to reconstitute the data deposit workflows by aligning project levels such that material, epistemic, and ethical 'thresholds' (criteria) obtain*. Faculty recursively do work to meet these passage points by performing this alignment work. In doing so, they not only ensure data deposit is *doable*, but also check data quality, prevent data loss, and ascertain data integrity.

Epistemic thresholds for data are mutually held criteria that judge the scientific validity, justification, and integrity of a dataset. Examples of epistemic thresholds include verifying data by reviewing images and dataset which students produce, e.g., P2, who reviewed western blots and P9, who partnered senior students with junior students to prevent biasing the quantification results of a hypothesis test. *Material thresholds* for data refers to the constraints which physical properties impose, and the steps taken to address these constraints. Examples of material thresholds include anticipating the decay of information technology and storage media and outmoded software. For example, P6 created paper and digital copies of experimental protocols and securely stored them on an intranet server accessible only by using university login

credentials and P3 upgrades equipment regularly and migrates data forward with the technology upgrades. *Ethical thresholds* for data are the normative commitments to which the scientific community adheres. For example, P6 invoked an ethical threshold when she prevented a student from publishing a paper based on data which had been selectively manipulated to make it appear as a positive experimental result.

Table 12: Threshold criteria for making data deposit *do-able*

Threshold	Description	Example
Epistemic	*Epistemic thresholds* for data are mutually held criteria that judge the scientific validity, justification, and integrity of datasets.	Verifying data by reviewing images and dataset which students produce, e.g., P2, who reviewed western blots and P9, who partnered senior students with junior students to prevent biasing the quantification results of a hypothesis test.
Material	*Material thresholds* for data refers to the constraints which physical properties impose, and the steps taken to address these constraints.	Anticipating the decay of information technology and storage media and outmoded software. For example, P6 created paper and digital copies of experimental protocols and securely stored them on an intranet server accessible only by using university login credentials and P3 upgrades equipment regularly and migrates data forward with the technology upgrades.
Ethical	*Ethical thresholds* for data are the normative commitments to which the scientific community adheres	P6 invoked an ethical threshold when she prevented a student from publishing a paper based on data which had been selectively manipulated to make it appear as a positive experimental result.

Thresholds are not mutually exclusive but are co-constitutive in that epistemic issues overlap with material and ethical ones (**Table 12**). For example, a common work task to deposit data is that faculty need to establish a chain of data provenance when submitting to a repository. In the case of P2, the data provenance chain involves tasks constituting a confluence of material, epistemic, and ethical issues. When preparing data for a publication, P2 instructs lab members to present their images and data to her for review (e.g., protein immunoblots) and data documentation to check if the timestamps ‘make sense’. The images and the documented timestamps provide a material representation of an epistemically important temporal piece of the justification of the data’s accuracy. The timestamps serve as physical evidence which P2 can use to judge whether the experiment was performed correctly, and the data can be used in the publication as initially intended. Similarly, images have material properties that researchers could alter with the intent to manipulate the data and misrepresent study findings.

The material, epistemic, and ethical dimensions of data intersect to constitute data quality and integrity. Data mediated through these thresholds imparts a higher probability for data accuracy and reliability, as well as long-term preservation. Yet, to be deposited, faculty must

make data satisfy thresholds by tinkering within and between with many different levels of work organization to achieve alignment. Epistemic, material, and ethical thresholds are instantiated by faculty scientists as evaluators of their data and repository guidelines. The thresholds are mechanisms located within and across the data deposit workflows that faculty work to achieve to ensure data quality, prevent data loss, and promote the integrity of data.

5.8 Conclusion

Public discourse and academic critique too often frame AI in genomics as mechanistic clockwork. Data work is often contrived in linear input/output (I/O) terms and repositories viewed as an unproblematic black box housing the I/O in a seamless flow where pristine data goes in and comes out. In contrast to the apparently straightforward process of submitting data to standardized and institutionalized repositories, we show that workflows are nonlinear. Depositing data to a research repository requires recursive efforts to make experimental work of data production and organization respond to concerns in multiple social worlds to conform to repository standards, secure resources, and publish papers. Our findings align with ample STS scholarship that has also shown the substantial labor of preparing data for being deposited, e.g. cleaning data (Darch et al., 2020b; Edwards, 2017; C. P. Lee et al., 2006; Leonelli, 2014a; Nadim, 2016; Plantin, 2019; Star & Griesemer, 1989).

By drawing from interviews with genomics faculty scientists in U.S. research institutions, this paper provides an empirical analysis of how scientists make data deposit *doable* through data articulation work. We adopt Fujimura's (1987) to foreground this work faculty scientists do to align three levels of work organization (experiment, laboratory, and social world) and two types of work (production tasks and articulation tasks) to construct a depositable dataset. We develop the concept of *datarticulations* to illuminate the practical negotiations faculty do to make data depositable. *Datarticulations* describe the recursive discretionary work scientists perform to reconstitute the data deposit workflows by aligning levels of work organization such that that material, epistemic, and ethical thresholds obtain. We argue this alignment work through thresholds not only ensures data deposit is *doable t*o enable AI, but also prevents data loss and ensures the quality and integrity of data.

Acknowledging and examining the articulation work inherent in genomics has implications for what data enters the repositories and the quality and integrity of that data. Big data approaches such as x-informatics rely on the aggregation of research data from

heterogeneous sources to build large datasets for machine learning, data mining, systems modeling, and other quantitative applications. For genomics and other x-informatics approaches to be effective, open research databases must accumulate as much data as possible. However, the assumption underlying that premise, as Leonelli (2014) points out, a “comprehensive” amount of data ensures scientific conclusions will be more robust to error and reliable. Yet, as both Kitamoto (2017) and Leonelli (2014) point out, the idea that more necessitates better quality science is misguided. Voluminous data is not equivalent to quality or comprehensiveness of data. Rather, what is needed is “a data-collecting strategy for collecting data [of] high quality" (Kitamoto, 2017). We show how faculty establish checkpoints at key places in data workflows to ensure data accuracy and reliability, enable access and interpretability of data. More broadly, examining the work of data deposit in situ disabuses us from AI imaginaries and helps us understand the human work behind data curation, work that has implications for AI.

If we are to understand the labor supporting CI-enabled science, we need to allocate attention to the ‘various and variably configured conditions of alignment of the many levels of work organization’ (Fujimura, 1987: p. 283). STS scholars can use the notion of *datarticulations* to understand how data is produced and science policymakers to design interventions for supporting faculty development in data management and genomics data workflows.

CHAPTER 6
STUDY 3

Researchers face a bewildering landscape of data management requirements, recommendations and regulations, without necessarily being able to access data management training or possessing a clear understanding of practical approaches that can assist in data management.

– Griffin et al. (2018)

6.1 Introduction

Research data management (RDM) has become increasingly overlaid with rules, policy, and expectations. This presence and pervasiveness of guidelines for RDM reflects the growing role of *institutions* in data management practices of academic faculty across many disciplines. *Institutions* are not only policies, but encompass the informal and formal rules, practices, and policies to form "durable social structures" (Scott, 2013) or "rules of the game" which shape human interactions (North, 1991). For example, the U.S. National Institutes of Health (NIH) and the National Science Foundation (NSF) created rules that Principal Investigators (PIs) must submit data management plans to receive federal funding (Diekema et al., 2014).

We see this rise in the *institutionalization* of RDM -- how these institutional rules become legitimate and taken-for-granted – because research data are seen as valuable assets that, if preserved, documented, and made accessible, can be standalone scholarly products and impact future research (Alperin et al., 2020). Institutions for research data management aim to support researcher efficacy and promote *long-term research data sustainability* – the reliability of the infrastructures supporting access to, and preservation of, data (Sands, 2017).

While institutions aim to support researchers, they can also complicate, or even inhibit, researchers and research. By adding new forms of work by institutionalizing RDM and deposit, which can require the researchers to make changes to existing practices and processes. For

example, compliance with RDM mandates is a new form of work, requiring researchers to do extra work in planning and writing a data management plan. Depositing data to repositories is a new form of work, requiring the extra effort, both intellectual and organizational: learning, gaining new expertise, organizing research, adjusting schedules, managing tenure and evaluation, deciding who should do the new work, adjusting to ever-emerging technologies (Akers & Doty, 2013; Diekema et al., 2014; Tenopir et al., 2014).

The extent to which these new forms of work in the form of institutional structures for RDM pervade researcher practices varies widely, depending on the discipline, methods, data (Akers & Doty, 2013). For example, 'big science' disciplines like genomics tend to have more institutional structures for RDM (Crowston & Qin, 2011). Social sciences and humanities less so (Akers & Doty, 2013). Disciplines with greater field-level institutionalization of RDM are associated with more mature RDM practices at the lab level, which, in turn, is associated with promoting long-term research data sustainability (Ankeny & Leonelli, 2015; Arias et al., 2015; Navale & McAuliffe, 2018).

Yet despite the increasing role of institutions in RDM and their critical implications for long-term research data sustainability, few studies have taken an institutional lens. The few studies that do draw from institutional perspectives in their analyses tend to be at the meso- or macro-level, e.g., a list of the types of institutional pressures (e.g., Diekema et al., 2014; Kim, 2013), a compilation of policy recommendations in specific disciplines (e.g., Bardach & Patashnik, 2019; Byrd et al., 2020; Corpas et al., 2018), or a historical view of field-level changes in data-sharing mandates (e.g., Arias et al., 2015; Hamid et al., 2009). As such, these do not engage the ongoing practices, attitudes, processes, and artifacts that *constitute* the work of data management happening daily in academic labs. Without such an analysis, scientists' work

that impacts *what* data is deposited and *whether* data is deposited at all – critical components of long-term research data sustainability— are inaccessible to scholarly analysis or practical support (e.g., science policy or system design). How does the institutionalization of RDM impact data practices? What institutional factors are associated with whether and what data gets deposited into repositories? Can disciplines with less institutionalized process for data deposit learn lessons from examining those who are more institutionalized? Addressing these questions will not only shed light on how we can better design RDM workflows for genomics and social science researchers, but also provide deeper insights into data policy in disciplines where cyberinfrastructure for RDM is less mature.

In this study, I examine the needs of scientists for RDM. Uncovering the characteristics and needs of scientists is the main goal, although examining the efficacy of data sharing mandates is not the main goal, I identify them if evidence appear along the way. This study informs policy by identifying institutional support successes and challenges in data management and deposit. The analysis also identifies workflow steps where policymakers can create interventions and direct resources. While research data management research includes many stages such as data collection and data analysis, in this paper, I focus on the deposit process because it offers the most assistance to scientists as prior research and our participants suggest.

6.1.1 Background

The debate on how to support data sharing in U.S. academic research has attracted much attention, given the massive investments in cyberinfrastructure (CI)-enabled science by U.S. federal agencies (*eScience* in the U.K.) (Atkins, 2003; Hey & Trefethen, 2005). In response, data management and sharing in many scientific fields has been increasingly institutionalized to varying extents, that is, formalized and standardized as an established part of the professional organization of many research fields (Crowston & Qin, 2011). A signpost of the

institutionalization of data sharing in CI-enabled science are open research data repositories. Examples of open research data repositories include the National Center for Biotechnology Information (NCBI) *GenBank* and *Gene Expression Omnibus* (GEO) and the *Inter-university Consortium for Social & Political Science Research* (ICPSR) and Harvard's *Dataverse*. Data repositories make data into an aggregated into a single body of information (e.g., with metadata). Making research data searchable and machine readable as a single body of information is seen as valuable by the scientific community because it enables scientists to perform computational manipulations on the data (e.g., using machine learning and artificial intelligence (AI) approaches) to do, e.g., comparative genomics and -omics.

As such, a key part of institutionalization and investment in CI-enabled science has been mandates to that require scientists to deposit data into repositories. For example, journal publishers require data sharing, federal agencies such as the U.S. National Science Foundation (NSF) mandated a data management plan, and scientific funding entities – private and public – require data sharing (Crowston & Qin, 2011; Kim, 2013). These mandates introduce a form of regulative institutional pressure that influences individual and organizations to deposit research data to repositories (Kim, 2013). *Institutional pressure* here refers to the environmental forces constituted of normative, regulative, and cultural cognitive pillars (Scott, 2013) which constrain and influence individual and organizational behavior. Individuals and organizations are driven to achieve *organizational legitimacy. In short,* institutional pressures are "the legitimizing means" that are derived from collective expectations (normative pillar), legal requirements (regulative pillar), and cultural comprehensibility (cultural-cognitive pillar)((Kim, 2013).

Yet despite the institutional interventions and studies of how faculty deposit data and which aim to facilitate high-quality data management, there is a gap between data policy and

local scientific practice. The gap is evidenced by well-known factors and obstacles to data management maturity (Crowston & Qin, 2011). As documented by numerous studies of data curation and management (e.g., Bratt et al., 2017; Larsen et al., 2014; Rehm et al., 2020), scientists' data management processes can often be inefficient, as they 'reinvent the wheel' every time they start a new project. Unstructured and ad hoc processes for RDM can also result in compromising the accuracy, reliability, and accessibility of data – among other attributes of quality (Herzog et al., 2007) – and thus, long-term data sustainability. This gap leads to the following research questions that drive the study.

6.1.2 Research Questions

This study is guided by a central question and several sub-questions. Given the perennial challenges associated with ensuring data sustainability, the critical role of scientists in making sure data is well-managed and deposited, and the rise of institutional structures and environment which aim to support them and mitigate data loss, the central question of this study is: How are research data management (RDM) and deposit in U.S. academic research groups shaped by an increasingly institutionalized environment for research data management? The central question leads to examining the extent to which institutionalization impacts long-term data sustainability. To guide our study, I focus on the following research questions (RQs) in the context of the research data management processes and focus on data deposit practices, processes, and artifacts:

- ***RQ1:*** What institutional factors are associated with articulation work during data management and deposit in the social sciences?
- ***RQ2:*** What are some of the impacts of articulation on long-term research data sustainability?

To address the RQs, a qualitative interview study was conducted to identify the extent of institutionalization of RDM – focusing on data deposit – and the possible associations with articulation work, as well as the possible impacts on long-term data sustainability. To do this, the study compares the RDM practices of genomics and social scientists and genomics U.S. academic research who deposit data.

The study uses a lens that combines *institutionalization* (W. R. Scott, 2013) and *articulation work* (Fujimura, 1987). In the discussion section, I discuss how these two come together using the theory of *due process in information systems.* On the whole, this paper builds on work examining the institutional environment for data management and deposit practices, and why they matter for how much and what kinds of "articulation work" they do (Fujimura, 1987), alignment work that has direct implications for the processes and outcomes associated with long-term research data sustainability (Bratt, Sharma, Erickson, *in prep*). Implied in the questions are issues of how we measure the extent of institutionalization of RDM. This will be addressed in the literature review, as well as the *Core Measures* section of the methodology. In the next section, I describe the background and review the literature on the institutionalization of RDM and data deposit, how it is measured, and studies of the impacts of institutionalization on long-term research data sustainability.

In this study, I use interviews and document analysis of research-active U.S. academic faculty who deposited data (n = 15) who deposit their research data to the Inter-university Consortium for Political and Social Research (ICPSR). I apply an "articulations" framework (developed in a prior study of data deposit in genomics) to examine the institutional factors associated with the articulation work of researchers. I use the framework as a sensitizing model to identify the institutional factors associated with articulation. The focus is on "articulation"

work because articulation can often be ad hoc, improvisational, and creative – but also associated with redundancy, lack of control, and inefficiency (Griffin et al., 2018; Lakhani et al., 2013). ICPSR is a useful context to understand how to support research data management because it is an emerging area where research data deposit to open data repositories is becoming more institutionalized. As such, it can be a rich site to develop policy for supporting RDM and data deposit across diverse research contexts and understanding the intersection of institutionalization and articulation more broadly.

The purpose of taking an *institutional perspective* is by surfacing the intersection of institutions and practices, enabling us to examine how institutional rules and norms shape researchers' practices, and practices impact institutional structures, in turn (Scott, 2013). The *articulation framework* is a useful analytic for revealing the work behind the data, such as documentation, metadata development, and preparing data for deposit to an open online repository. In combination, the two theoretical perspectives surface the factors associated with articulation work involved in preparing data for deposit to an open research data repository.

In this paper, I first provide background on ICPSR and data deposit and review related work on institutional factors shaping research data management and deposit. I describe the study methods, including the core measures used to operational 'institutionalization' and 'articulation.' I find scientists are incorporating field-level policy (e.g., NSF/NIH data management mandates) and norms (e.g., I argue these empirical data show developments that are evidence of *institutionalization*, that is, the process whereby practices because taken-for-granted and legitimate (Scott, 2013). The less-institutionalized contexts are developing local structures such as templates for research data management.

I discuss the potential ways that less-institutionalized contexts can learn from the more institutionally mature disciplines to make their data findable, accessible, interoperable, and reproducible (FAIR) (Wilkinson et al., 2016). I discuss the implications of the findings for long-term data sustainability and issues of institutionalizing data management and deposit, and develop the concept of *articulating institutionalization*, drawing from the *due process in information systems* work (Gerson & Star, 1986).

The chapter concludes with a summary and recommendations for science policy and scientists' practices and processes and suggest future research directions. With these understandings, researchers can better isolate variables and develop models to support data deposit. Science policymakers can plan interventions and allocate resources to prevent data loss to for long-term research data sustainability and professional organizations can assist with faculty development. More broadly, examining the work of data deposit disabuses us from the grand vision of cyberinfrastructures by helping understand the work involved in data management and deposit in which scientists engage to meet emerging institutional demands.

6.2 Related Work

In this section, I describe research data management and deposit and how it has been institutionalized over time across scientific fields, giving the broader context of mandates and policy, as well as normative and cultural institutional forces. In this, I focus on research data repositories, a signpost of institutionalization of RDM and deposit and key institutional force. I then discuss literature investigating the impacts of the institutionalization of RDM on long-term sustainability, including theories of articulation work, which describe the work of managing and depositing data in contexts with many institutional structures for RDM, and those with low levels of institutional structures for RDM. Next, I review the existing measures of the

institutionalization of RDM (e.g., RDM assessment tools). I elaborate on the context of the study: depositing data to ICPSR data deposit requirements Finally, I synthesize the literature to show few studies existing measures, indicating a need to connect (neo)institutional theory to assessment frameworks of RDM maturity (e.g., the *CMM for RDM*), a methodological gap addressed in part in this study.

6.2.1 Institutionalization of RDM

Institutionalization refers to when the ongoing processes that lead to – and maintain – beliefs and practices as natural and taken-for-granted. Institutional logics are a key part of the institutionalization process, as they are the ways that systems of beliefs inform actors decisions (Thornton & Ocasio, 2008)[9]. A mainstream theory of institutions is that of Scott (2013), which has evolved into what is referred to as *neo-institutional theory* is Scott's "three pillars of institutions" (2013). The three pillars are a metaphor for the categories that construct institutions. Each pillar has a basis of compliance, order, and legitimacy; as well as mechanisms, logic, and indicators (**Table 13**).

Table 13: The pillars of institutions, and their basis of compliance, order, and legitimacy; and their mechanisms, logic, and indicators. Adapted from (Scott, 2008).

	Regulative	**Normative**	**Cultural-cognitive**
Basis of compliance	Expedience	Social obligation	Taken-for-grantedness Shared understanding
Basis of order	Regulative rules	Binding expectations	Constitutive schema
Mechanisms	Coercive	Normative	Memetic
Logic	Instrumentality	Appropriateness	Orthodoxy
Indicators	Rules, Laws, Sanctions	Certification, accreditation	Common beliefs Shared logics of action
Affect	Fear, Guilt/Innocence	Shame/Honor	Uncertainty/Confusion
Basis of legitimacy	Legally sanctioned	Morally governed	Comprehensible, recognizable, culturally supported

[9] For further details of institutional pillars, institutionalization, and related concepts refer to Ch. 2 – Literature Review in this document.

The indicators are particularly useful in informing the measures of the institutionalization of RDM. For example, an indicator of the institutionalization of data deposit from the perspective of the *regulative* pillar is the legal component of the GenBank repository that enforces an embargo of datasets until a determined date of release (Benson et al., 2017).

Institutional pressures constrain and enable behavior. As such, they have implications for how people act. In the case of data deposit, the institutional pressures to deposit data (to open research data repositories) influence how scientists act – specifically how they manage and share data (S. Kowalczyk & Shankar, 2011; Sandusky et al., 2021; Shankar & Eschenfelder, 2017; Tenopir et al., 2014). Studies of the *regulative* institutional pressure, e.g., NSF/NIH mandates to deposit data, show how these policies have led scientists to scientists strategically withhold some data (e.g., Hrynaszkiewicz et al., 2020), but also that funding mandates and state-sponsored infrastructures (e.g., Priego et al., 2022), among other legal and policy, shape data sharing ethics and practices (e.g., Fiesler et al., 2020).

Normative institutional pressures like the Open Science movement, studies suggest that the moral imperative to share data are associated with the extent to which scientists put data on preprint servers (e.g., arXive)(e.g., Thursby et al., 2018). Moreover, studies focusing on the *cultural-cognitive* pressures, studies find, impact how collectively-understood and "orthodox" the heuristics are for managing and sharing data (e.g., Corpas et al., 2018; Kim, 2013).

Empirical work has suggested institutional factors associated with data sharing ('sharing' includes data deposit). Kim (2013) found three factors in his work on the individual and institutional factors influencing researchers in STEM disciplines to share data: disciplinary associations, journal publishers, and funding agencies. Kim also adds that each of these can be considered within the institutional framework of Scott's three institutional pillars (regulative,

cultural-cognitive, and normative) (2013). Regulative pressure comes from funding agencies and journals, while normative pressure comes from the discipline. In addition, Kim identifies *metadata* and *data repositories* as possible sources of discipline-level institutional pressure to share data (Kim, 2013, p. 202). Disciplinary associations such as professional conferences (e.g., the American Psychological Association) and accreditation through education and training of students Kim (2013) suggests are potential institutional pressures that factors into researchers' data sharing activities. However, as of 2013, funding agencies did not exert pressure on researchers to share their data, as Kim (2013) writes: "This research suggests that funding agencies need to enforce their data sharing policies after awarding grants… regulative pressure currently exhibited by funding agencies does not have a significant effect on scientists' data sharing behaviors across different disciplines" (*ibid*, p. 217).

Funders to incentivize data sharing through mandates (e.g., the NIH and NSF data management plan and sharing mandates) (Arias et al., 2015; Byrd et al., 2020; Corpas et al., 2018). It is more common for federal funding sources to require researchers to include requirements for researchers to share their data, but in some disciplines, foundations also have begun to require data deposit to a repository. However, it is still relatively rare for foundations to be major coercive pressure to share data via repository (Borgman et al., 2019; Khan et al., 2020).

Journal publishers have also been found to be associated with data sharing. Journals in each discipline need to require their authors to share data for their published articles (Hrynaszkiewicz et al., 2020; Kim & Adler, 2015). Publishers can require authors to deposit data to open research data repositories before publication if no related repositories exist, and in the case where no relevant repository exists for their data, researchers are required to provide data to requesters (Kim & Adler, 2015). Studies also find normative pressures do influence the extent of

data sharing in some fields, that is, the “social and moral obligations” researchers experience that lead to data deposit (Kellogg et al., 2006).

Education and training in each discipline can help scientists develop similar disciplinary norms about data sharing in the form of scientific ethics (DiMaggio et al. 1983), and professional associations and accreditation agencies in scientific communities can actually exert normative pressures with regards to data sharing (Grewal et al. 2002). Each scientific community can develop their norms of data sharing through education and training that are supported by their professional associations and accreditation agencies. Cox & Pinfield (2013) examined the role of funder mandates in organizational perceptions towards RDM, finding that funder mandates are key drivers for positive shifts in perceptions of RDM.

More recently, scholars have drawn from information systems (IS) and economic theory to explain the how scientific norms (e.g., Mertonian norms) and incentives/disincentives shape whether scientists share data (Murray & O’Mahony, 2007; Stephan, 2012a). For example, Priego et al. (2022) explain the “puzzle of sharing scientific data” using the sociology of science theory epistemic cultures and action theory. In the paper, the authors examine major barriers and enablers of sharing, with a focus on the cultures of work in specific fields focusing on the Human Genome Project in molecular biology (MB) and high-energy physics (HEP). Priego et al. (2022) develop a framework of the mechanisms that ease tensions between “the community epistemic norms and the individual costs and benefits of data sharing” between MB and HEP (**Figure 10**).

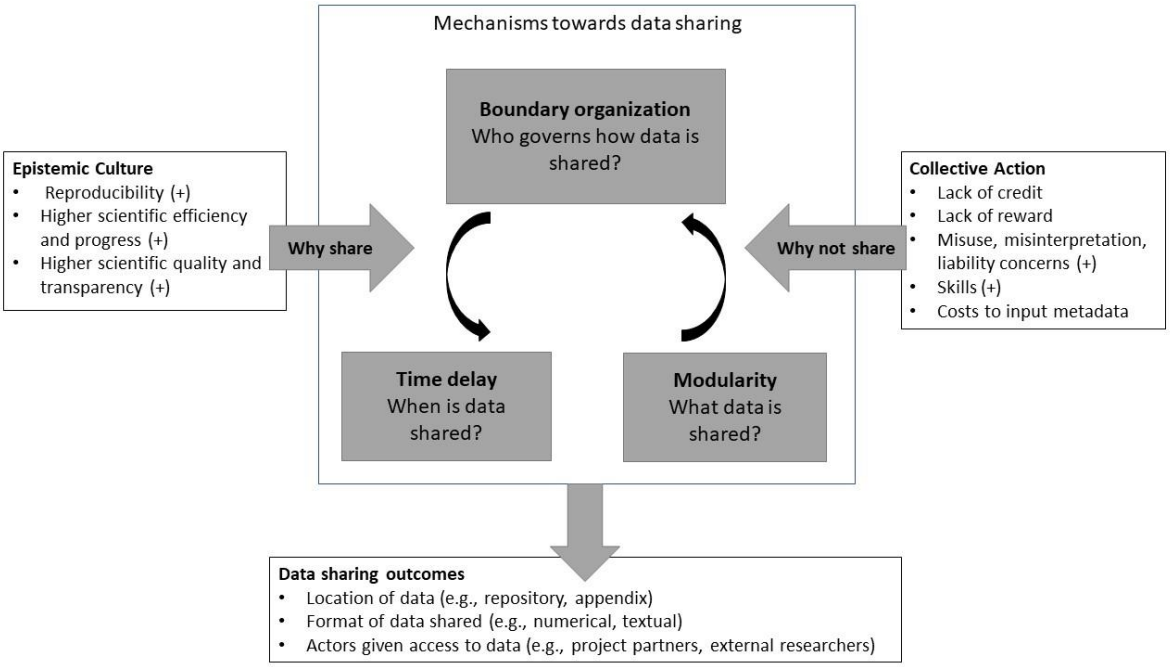

Figure 10: Framework of the "mechanisms" that enable researchers in Molecular Biology and High Energy Physics to share data. Adapted from: (Pujol Priego et al., 2022).

To operationalize the institutionalization of RDM, the Capability Maturity Model (CMM) for Research Data Management (RDM) is a model describing the institutionalization of RDM in terms of maturity. That is, the CMM for RDM was created as a tool which assess the extent to data management, as a process, is "mature," essentially a type of institutionalization. In fact, the CMM for RDM defines institutionalization of a process. For example, the "institutionalization of a managed process" is defined as when an organization has "policies for planning and performing the process, a plan is established and maintained, resources are provided, responsibility is assigned, people are trained, [and] work products are controlled…" (Crowston & Qin, 2011, p. 17). There are five key process areas which (Crowston & Qin, 2011) define for research data management: 1) data management in general; 2) data acquisition, processing and quality assurance; 3) data description and representation; 4) data dissemination; and 5) repository services and preservation. Each key process area is further divided into a number of sub-areas.

The description of these sub-areas includes definition of key concepts, rationale/importance, examples, and recommended practice.

Existing studies offer insight into the kinds of institutional pressures that shape faculty research data sharing. They provide practical recommendations for organizations, funding agencies, policymakers, and other organizations to inform data policy. They also provide an analytic framework focusing on institutional pressures and incentive structures that allow further research to understand the relationship of institutions and data management. However, the most pertinent studies of institutions and practices are somewhat outdated, before the recent and rapidly unfolding policy developments in policy mandates, normative pressures, and the accelerating technological innovations for RDM (e.g., Blockchain for RDM workflows (Chen et al., 2018), Jupyter Notebooks (Kluyver et al., 2016), REDCap (Harris et al., 2009) and cultural developments (e.g., the open science movement (Randles et al., 2017), open access (Den Besten et al., 2009). Often, these studies take the perspective of library practitioners who tend to focus on the end of the research data lifecycle, such as on preservation (Navale & McAuliffe, 2018), archival techniques (Borgman et al., 2019), and researcher desires for data management and digital scholarship services (Akers & Doty, 2013; Joo & Peters, 2020; Kollen et al., 2017).

These studies have begun to take an explicitly infrastructural, cultural, and institutional perspective, leading to valuable insights into the intersection of institutions and practices. For example, Darch et al. (2020) focus on how different library cultures use existing institutional resources and interpret data policy to enable astronomy data archiving. A study of the role of institutions in data management and deposit is one that exhorts the institutionalization transparency in social science, Freese & King (2018) implicitly operationalize the institutionalization of research process transparency as the rules, policies, actions, and resources

of 5 institutional actors and collection of actors. They imply that institutional actors are journals, reviewers, professional organizations, teachers and mentors, universities and departments, funding and data sources (Freese & King, 2018). In this description, they refer to *routines* and *norms*, implicitly invoking institutional concepts to make their case without explicitly drawing from institutional (or neo-institutional) theory. The authors present specific actions which various institutional actors involved in research practice in the social sciences can take to promote transparency in knowledge production (Freese & King, 2018). Moreover, recent studies have taken a "practice turn" to uncover the decisions "behind and beyond the datum" including the subtle, complex practices of managing and data sharing (Neang et al., 2020).

Yet an institutional perspective that focuses on the intersection of institutional factors and the everyday research data practices in the social sciences has yet to be examined. What theories can help us understand the intersection of institutions and RDM that enable deposit? Prior work has suggested "articulation" is one framework useful to explain what makes data deposit "do-able" (Bratt, Sharma, & Erickson, *in prep*). In the next section, we turn to the "articulation" theories and conceptual frameworks.

6.2.2 Articulation Work in RDM

"Articulation" and "articulation work" are concepts with multiple meaning, depending on the author. Articulation has its origins in medical work, as developed by Anselm Strauss (Corbin & Strauss, 1993; Star & Strauss, 1999; Strauss, 1985, 1988). It has been applied to other contexts, including software development (e.g., Boden et al., 2008), knowledge management in the enterprise (e.g., Suchman, 1995), and often in critical scholarship on "invisible" or undervalued work , that is, work that is often overlook or lacks credit-attribution (Schmidt, 2002, 2016; L. A. Suchman, 1996; Walker et al., 2019). A theory especially pertinent to

questions of research data management and deposit is by Joan Fujimura (Fujimura, 1987). Fujimura (1987) develops the concept of articulation work to explain what makes a scientific problem "do-able," that is, selected and carried out after being determined to be a feasible project. Elaborating on Strauss' articulation concept, Joan Fujimura (1987) describes articulation as the alignment of multiple levels of work organization. That is, articulation is the work of making sure the various levels align – e.g., when stakeholders at the social world level need an outcome from the laboratory or experiment level (see **Figure 9**).

"Articulation work" has implications for the efficiency of data management workflows, successful data deposit, and data quality – all which ultimately impact long-term research data sustainability (Lakhani et al., 2013). "Articulating" is often ad hoc, creative, and improvisational, which has implications for how data is managed, documented, and preserved. Articulation work can often be ad hoc and improvisational (Bossen et al., 2019), such as responding to unexpected events or interruptions in data collection. As a result of work not being documented, redundancy can occur, inhibiting workflow efficiency (Bishop et al., 2020). Articulation work also often involves creativity (Schmidt, 2002; Strauss, 1985), e.g., dealing with unexpected situations in nursing work (Bowker et al., 2001; Bowker & Star, 2000; Schmidt, 2016). Creative solutions, such as in dealing with data by technicians (Plantin, 2019; Shapin, 1989), is no predefined set of steps to take but requires drawing from resources "at hand" (Nardi & O'Day, 1999).

Despite the valuable contributions of this work to our understanding of data deposit, we still lack substantive theory to address how data is deposited. Building from this work, this study addresses the gap of the few studies which specifically address the intersection of institutionalization and articulation, in the specific context of RDM and deposit.

6.3 Methods

The study takes a qualitative approach by applying the "articulation" framework to a broader population to surface the institutional factors associated with articulation work during research data management (RDM) and deposit to an open research data repository. The data consists of in-depth, semi-structured interviews with U.S. faculty at academic institutions who deposited their research data to the Inter-university Consortia for Social and Political Science Research (ICSPR) within the last 5 years (2017- 2021). Recruitment criteria and justification for the selection of participants is detailed further in the *Data Collection* section below.

In addition, I collected documents related to the dataset documentation associated with their deposits as well as related research data management documents, which were requested and provided by the faculty (e.g., a lab handbook). Faculty researchers who deposit to ICSPR were selected because they represent a sample population who are in a field with a growing, but still developing, institutionalization of data management and deposit. This sample population allows us to test the articulation framework and hold other relevant factors equal to identify institutional factors that shape faculty data management and deposit work.

The data analysis approach includes inductive and deductive elements. First, open coding occurred to allow for themes to emerge and then the "articulation" work was applied. To analyze the extent the institutionalization of RDM, and how RDM manifests across the research data lifecycle and to enable data deposit, a model from the literature was selected – the Capability Maturity Model for Research Data Management (CMM for RDM) (Crowston & Qin, 2011) – because it is relevant to the context of research data management, and an authoritative model in the literature often used to assess the maturity of RDM in academic labs. More details and justification for these approaches are provided in the following sections (*Data Collection*, *Data Analysis*, and *Operational Measures*).

6.3.1 Data Collection

This study follows a semi-structured interview study approach (Creswell et al., 2007). The semi-structured interview study approach was selected for its utility in investigating a specific issue (i.e., the impacts of data institutionalization on faculty practices and implications for long-term research data sustainability) using interviews to best understand the research phenomenon (Creswell et al., 2007). The research design is an approach which applies the "*data articulation"* framework developed in Bratt, Sharma, & Erickson (*in prep*).

Documents were also collected. The participants were asked to provide research data management documents, as well as the record of the dataset(s) they deposited were collected by saving the website as PDFs as well as the hyperlinks to re-visit during analysis. The purpose of the data management documents was to supplement the interviews and for additional information about how the data management and analysis were ostensibly written up.

6.3.1.1 Interviews

The interviews were conducted via zoom video conference in September – November 2021. The interview method was selected to elicits data, usually verbal data in the form of a transcript. A protocol is created with questions and prompts so the interviewer interacts face-to-face or through synchronous video conferencing in an alternating series of questions and responses. Interviews are audio-recorded and securing stored before transcription into a document for analysis. The advantage of the interview methods is their richness as a collection instrument, probing for unexpected topics and deeper understanding about the topic at hand.

The semi-structured research interview instrument used here is a questionnaire designed to focus on the perceptions, behaviors, and attitudes of researchers. The first section of questions is to establish a background and professional history of the interviewee. Critical incident-inspired questions to elicit stories and experiences from participants are designed into the section on

interactions with online systems, specifically, data management practices and best practices. Interviewees are asked to provide chronological, sequential description of the conceptual and computation steps they take to initiative, execute, and complete data tasks.

In-depth interviews ask interviewees to provide illustrative examples of their experiences as well as personal reflections about changes over the course of their professionalism experiences in data-intensive research projects, changes in knowledge and information sharing and the research assessment such as personal expectations for their work and the formal processes in academic research for judging quality, impact, progress, and overall ‘good work.’ Speculation about the beliefs, behaviors, attitudes, and perceptions of other researchers are not specifically encouraged but are admitted as part of the emergent process of semi-structured interview elicitation.

To ensure a consistent representation of data deposit experiences, we recruited research-active faculty at U.S. academic institutions (R1=14, R2=1). Study eligibility was limited to participants who deposited data to ICPSR within the last 5 years (2017-2021) and whose data was qualitative. The purpose for selecting faculty who employed a qualitative approach and submitted qualitative data was to control for the known effects of methodological paradigm (i.e., qualitative versus quantitative) on data deposit practices. For example, in a study of faculty data management practices, Whitmire et al. (2015) suggested a strong correlation between quantitative data and the ease of sharing to an institutional repository. A higher rate of deposit compared to humanities were attributed to the ease of upload given existing metadata. Similarly, Van Tuyl et al.'s study (2015) of disciplinary differences documented the effects of the type of methodological approach and associated data type (i.e., qualitative /quantitative) on data management and deposit practices in STEM contexts. Thus, faculty were selected whose data

deposit included qualitative data because we assume, as prior work suggests, the methodological orientation (i.e., qualitative vs. quantitative) effects data deposit and management practices.

The purpose of selecting for only scientists who deposited qualitative datasets to reduce a known source of variability of data deposit practices (Antes et al., 2018; Mannheimer et al., 2019; Mozersky et al., 2021). The *data type* – qualitative versus quantitative – is a known source of variability that impacts the kinds of articulation work needed to successfully deposit data in response to institutional pressures to deposit data (i.e., institutionalization). Not controlling for the data type would confound the study. Qualitative data faces notorious challenges to being deposit relative to quantitative data. While it is out of the scope of this study to detail exhaustively the relevant differences between qualitative and quantitative data work with respect to articulation during data deposit, there are two factors relevant to articulation work associated with data type worth relevant to data deposit.

First, scientists who submit qualitative data must add metadata and other descriptive elements to their dataset submission to ICPSR. Note that qualitative data may include not only interview transcripts, but also photos, audio, and other forms that often need cleaning, description, and anonymization. Granted, quantitative data largely are not self-described – even if the columns name the variable. Researchers still need to add metadata to quantitative studies.

Second, scientists do not deposit qualitative data because it may breach confidentiality, even if their participants consent and they anonymized the data, there remains a risk of participants' identity revelation. Scientists perform articulation work to ensure data is made confidential for deposit (e.g., by anonymizing transcripts). Quantitative data are less vulnerable to confidentiality concerns because they are aggregated. Quantitative data require relatively less articulation work because they do not require anonymization, nor following-up with participants

for consent to deposit the data, reviewing transcripts for errors, and iteratively emailing with ICSPR staff for approval of the anonymized qualitative data. In sum, the differences concern central aspects of the study research questions: 1) *articulation* and 2) *institutionalization*.

Recruitment proceeded by solicitation via email and by phone. Participants were informed that interview would last approximately 60 minutes. The sample size was 15 faculty. The sample size that exceeds the approximate recommended sample size for interview data (n = 11-20, depending on the representative sampling philosophy of the research paradigm and methods) (Creswell et al., 2007). In addition, thematic saturation was becoming apparent after the 15 faculty (Creswell et al., 2007) (**Table 14**).

Table 14: Study 3 Participant Demographics

ID	Gender	Position/Title	Discipline
P1	Non-binary	Assistant Professor	Sociology
P2	Female	Associate Professor	Epidemiology
P3	Female	Clinical Assistant Professor	Learning health sciences
P4	Male	Professor	Criminology
P5	Male	Associate Professor	Criminology
P6	Female	Associate Professor	Health sciences
P7	Male	Professor Emeritus	Social epidemiologist
P8	Female	Associate Professor	Public policy
P9	Female	Associate Professor	Adolescent development
P10	Female	Assistant Professor	Health services and policy
P11	Female	Assistant Professor	Epidemiology
P12	Female	Professor	Survey methodology
P13	Female	Adjunct Associate Professor	Criminology
P14	Female	Assistant Professor	Health services
P15	Female	Associate Research Scientist	Social work

Participants were selected as those who deposited to ICPSR successfully because I needed successful cases to identify the articulation work associated with successful deposit. To identify that these "articulations" were indeed not accidental or random, or activities that occurred regardless of depositing data or not, but indeed associated with data deposit, I asked the scientists to report about the necessary actions to deposit data, not the actions incidental to the scientific research process. Yet the reader may notice the sample only includes successful data deposit, implying that point instances when data were not deposited may have indeed been incidental to the scientific research process.

The interview protocol was designed to ask scientists also about the instances when data were *not* successfully deposited to address this potential concern, as well. The inclusion of successful and unsuccessful data deposit allowed for a within-subject comparison and a between-subject comparison of successful data deposit and unsuccessful deposit – or when data were not constituted *as* datasets. While out of the scope of the study to address the constitution of data *as* datasets for deposit, see the *Discussion* for brief commentary.

6.3.1.2 Documents

Documents were collected from participants. The criteria for collection were broadly defined as those related to research data management and deposit; that is, the data management plan required by a funder were requested from the faculty member to show the scientists' ostensible plan for research data management. Informal documents were collected. If the document was mentioned in the interview as supporting the data management and deposit process, e.g., interviewer protocols. A total of 28 documents were collected. They included formal research data management planning document, as well as internal documents such as protocols for interviews and templates for qualitative data analysis.

The datasets the participants submitted were a variety of data types, were supported by researcher funders, collection methods, dates of collection, data use restrictions. Their associated outputs and usage metrics (e.g., downloads). The datasets record details to which participants referred for this study are shown in **Table 15.**

ID	Datatype	Funding	Collection method	Collection date	Down-loads	Data-related publications
P1	audio: sound data; survey data; text	University Internal Funding: Graduate Research Assistant stipend	face-to-face interview; telephone interview	2019	18	1
P2	text; interview data	US Department of Health and Human Services; Agency for Healthcare Research and Quality; DOD; Bayer, Novartis, and Pfizer; Janssen/J&J, Novo Nordisk, Regeneron, Sanofi.	focus group; individual interviews; face-to-face interview; telephone interview	2015- 2016	16	1
P3	text; interview data	Completed as a requirement for a PhD in Palliative Care	face-to-face interview; In-depth qualitative interviews with Masters-trained ward social workers	2014-2015	6	1
P4, P5	text; survey data	United States Department of Justice. Office of Justice Programs. National Institute of Justice	face-to-face interview	2016-2018	255	22
P6	text	National Institute of General Medical Science, National Institute on Minority Health and Health Disparities, Louisiana Cancer Research Consortium	face-to-face interviews; focus groups; demographic survey questions	2015-2016	22	3
P7	Survey data	United States Department of Health and Human Services, NIH, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Office of Behavioral and Social Sciences Research, Office of Research on Women's Health	mail, telephone interview, web-based survey	2016, 2017-2018	7,225	17
P8	Survey data	National Endowment for the Arts	Telephone interview, web-based survey	2018	868	4
P9	administrative records data; text; survey data	United States Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention	face-to-face interview, on-site questionnaire, telephone interview, web-based survey	2013-2015	64	1
P10	Text	Substance Abuse and Mental Health Service Administration	Telephone interview	2018-2020	16	5+[10]

[10] The associated publications showing in the ICSPR dataset was “0”; However, there were more than 5 associated with medical-assisted opioid treatment, the subject of the study, on the participants’ Google Scholar profile.

		(SAMHSA), Health Resources & Services Administration				
P11	Event/transaction data; observational data; survey data; administrative records	Chicago Community Trust	computer-assisted personal interview (CAPI), face-to-face interview	2015-2016	1,266	6
P12	Text	National Science Foundation. Directorate for Social, Behavioral, and Economic Sciences	Survey; Computer-assisted telephone interview (CATI)	2013	426	3+ (see footnote 10)
P13	Survey data	United States Department of Justice. Office of Justice Programs. National Institute of Justice	Face-to-face interview	2018-2019	42	1
P14	text	Boston University Clinical and Translational Science Institute	face-to-face interview, telephone interview	2018	27	2
P15	Survey data	United States Department of Health and Human Services. National Institutes of Health. National Institute on Drug Abuse	on-site questionnaire, telephone interview	2009-2013	115	0 (see footnote 10)

Table 15: Characteristics of the datasets submitted to ICPSR by study participants.

Included in **Table 15** are the funding source(s), data type(s), method of data collection, the number of associated publications, dates of data collection, number of downloads since initial deposit, restriction status, and primary purpose for depositing. In addition, the record for the faculty's data deposit to ICPSR (or other data repository) was saved as a PDF in a local folder. The record was used to provide information about the type of data they submitted, the data, and the associated metadata. The purpose was to provide further documentation that inform insights into the institutionalization of RDM and deposit in ICSPR.

6.3.2 Data Analysis

The data articulation framework is thus tested and elaborated by using the constant comparative method, drawing from both deductive and inductive coding techniques enabled by this method. The *constant comparative* approach to qualitative data analysis is an approach developed by Barney Glaser (1965) to generate and elaborate theory more systematically The

Therefore, this number is corrected by the author to reflect the productivity of the dataset on other scholarly product aggregators (i.e., Google Scholar).

constant comparative method is useful for testing and elaborating theory. It combines two approaches to qualitative data analysis: (1) deductive analysis approach (using explicit coding and analytic procuedures), and (2) inductive approach. Inductive is better at "generating theoretical ideas – new concepts and their properties" by applying a more heuritic, less systematic way, that is, "….constanting redefining and reintegrating his [sic] theoretical notions as he reviews his [sic] material…" (Glaser, 1965, p. 437), but less systematic than deductive, since "the analyst merely inspects his data for new properties of his theoretical categories and writes memos on these properties" (Glaser, 1965, p. 437). Constant comparative coding, then, integrates these two approaches to enable a systematic, inductive-deductive approach by drawing from the explicit, systematic coding techniques of the first approach as well as the constant reintegration of notions and concepts of the second.

The audio from the interviews was transcribed using a semi-automated transcription software (*Rev.com*). The interviews transcript data was combined with the source secondary documents (e.g., faculty website snapshots) and observational report data within the *nVivo* software (version 12), musing the multiple data sources feature. The first pass of the interview data was analyzed using the qualitative data coding software NVivo. This part of the data analysis method was drawn from grounded theory as an inductive approach, using iterative memos and axial coding to identify themes and sub-themes in the data. In grounded theory, *axial coding* is a process of relating the codes from the text to each other. According to Strauss & Corbin (1994), these relationships create a "coding 'paradigm" that encompasses themes (codes and categories). The codes in the paradigm relate to: (1) the research phenomenon of interest; (2) the context and conditions related to the phenomenon under study (i.e., the structural, intervening, and/or causal conditions); (3) the actions taken and consequences of the actions

(referred to as the "interactional strategies") involved in the phenomenon (Charmaz, 2006; Kelle, 2005). The codes and themes of the exploratory study are combined to inductively analyze the types of invisible data practices that appear and the factors contributing to their invisibility.

In a second pass of the data, a deductive qualitative data analysis approach will be taken to test the research framework in the molecular biology and biochemistry. In the deductive pass, the proposed research framework will be tested to map the consolidated conceptual activity to the emergent data practices. If the results of the data analysis surface new invisible data practices, the research framework will be modified, and the molecular biology data will be reassessed in juxtaposition with the modified and refined framework as a validation step.

6.3.3 Operational Measures

Empirical studies of faculty data management and data deposit practices do not explicitly use (neo)institutional theory to classify genomics as reflecting "higher-institutionalization" of research data management and social sciences as "lower-institutionalization" of research data management. However, it can be observed that the fields have the distinctive characteristics of institutionalization. The empirical findings of these studies show that the three institutional pillars (regulative, normative, and cultural cognitive) are at work in genomics to legitimize research data management and naturalize specific process, policies, and norms for data management as 'taken-for-granted' (Freese & King, 2018).

The phenomena measured in this study are a) *institutionalization of RDM* and b) *articulation work in RDM*. These concepts are defined on a spectrum and can be measured using existing frameworks. Specifically, the *institutionalization of RDM* is measured by selecting core measures draw from the capability maturity model for research data management (CMM for RDM). The core measures for *articulation work in RDM* are measured by indicators drawn from an adaptation of Joan Fujimura's (1987) work on articulating alignment to make scientific

problems 'do-able.' Institutionalization can occur at multiple levels of analysis – and differs across different models and theories of institutionalization (*Cf* (Barley & Tolbert, 1997; Crowston & Qin, 2011; Scott, 2013). The study focus is placed on data management and deposit. The corresponding unit of analysis is the data practices of faculty. The *research group* was selected as the level of focus because it is the sphere of control which faculty make decisions that impact data management and deposit, e.g., setting up data workflows and supervising students.

6.3.3.1 Measures for Institutionalization of RDM

In this study, the CMM for RDM is used as a starting point for an operational measure of institutionalization of RDM and data deposit. A full assessment using the full CMM for RDM is out of the scope of this study[11]; therefore, a subset of core measures was drawn from the model to represent the institutionalization of data management and deposit (**Figure 11**).

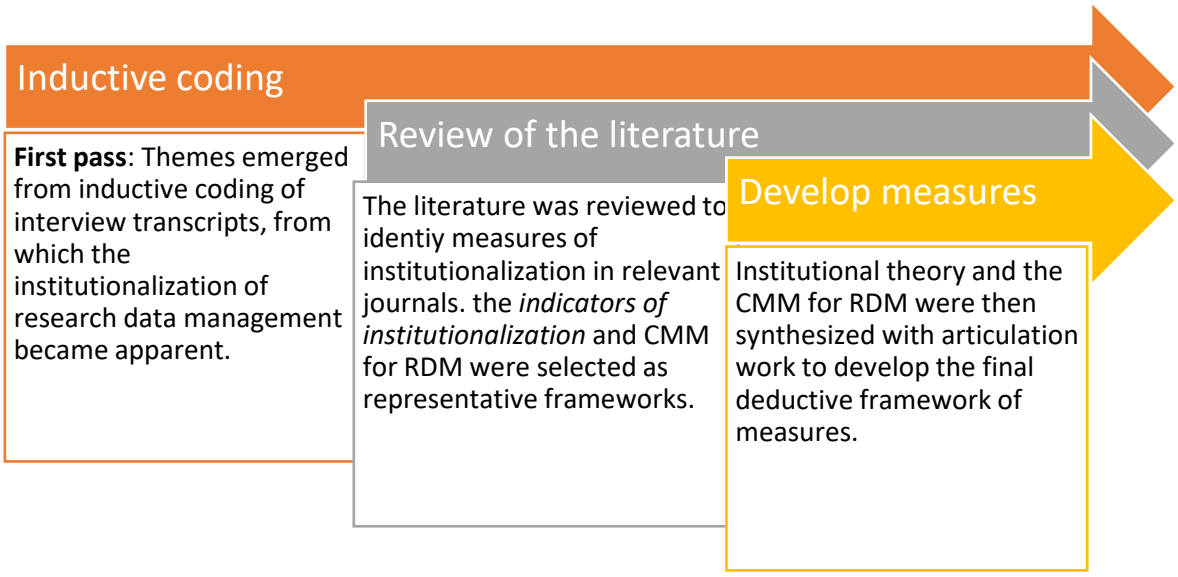

Figure 11: Process and method for selecting measures of institutionalization of RDM

[11] Amounting to analyzing the data for over 400 measures (4 areas indicating of maturity x ~4 items each x 5 levels of maturity x 5 process areas. There are 378 items total.

These core measures were selected to reflect the primary aspects of institutionalization of data management and deposit at the research group level. An initial content analysis and inductive coding of the interview transcripts was performed, allowing for the themes of how faculty managed and deposited their data to ICSPR. Then, the literature was reviewed to identify existing measures. Institutional theory and the CMM for RDM were used to provide a foundational view (institutional theory) and a context-specific set of measures.

Next, the CMM for RDM including the rubric for RDM maturity was reviewed and all relevant process areas were selected. The resulting subset of core measures were selected from the generic process areas, with a focus on the processes most relevant to data deposit (e.g., *Data Management in General* and *Data Sharing and Dissemination, Repository Services*), focusing on the RDM work required to prepare data for data deposit. Then, a second pass of the data using the operational measures was performed, using the resultant measures. These core measures capture key aspects of the institutionalization are RDM, but they are an initial starting point. Future work will further develop the core measures, e.g., using a survey approach, to identify key indicators of the institutionalization of RDM.

6.3.3.2 Measures for Articulation Work and Institutionalization in RDM

To measure articulation work and institutionalization of RDM, it was necessary to identify a preliminary deductive framework that brings Fujimura's (1986) measures and levels of articulation and the measures and levels of Scott's (2008) institutionalization. The core measures of the articulation work done to manage RDM institutionalization are shown in **Table 16.** There are three analytic levels of institutionalization in Fujimura's (1987) model of articulation: the

experiment level, the research group/lab level, and social world level[12]. These levels (of articulation) are compatible with the levels outlined in Scott (2013) on institutional theory and the CMM for RDM (Crowston & Qin, 2011).

Table 16: Core measures of institutionalization and articulation of research data management and deposit

Analytic Levels of Articulation Framework	**Core Measures of the Institutionalization of Research Data Management (RDM)**
Level 1: Experiment Level	• Presence/absence of documentation • Presence/absence of structured workflows • Extent of rules about validity in research design **Focus:** Activities, artifacts, and processes reported by faculty that reveal how researchers organize data and manage workflows during early stages of data management (e.g., collection, analysis) to prep data for deposit.
Level 2: Research group/lab Level	• Extent of standardization of data sharing within the research group • Presence/absence of newcomer orientations protocols • Extent of file naming conventions • Extent of agreement on whether to deposit data **Focus:** Activities, artifacts, and processes reported by faculty that show how researchers coordinate with research group during the data deposit process.
Level 3: Social world Level	• Presence/absence of policy (e.g., federal) • Extent of disciplinary rules that influence researcher identity **Focus** Activities, artifacts, and processes reported by faculty involved in how researchers interact with data repositories, professional associations, funders, and/or publishers.

The articulation work framing to helps us to understand data deposit because it constructs data deposit as a process of reconciling institutional mandates and local practices to make data 'do-able,' that is, suitable for deposit to an open online research data repository. Articulation work is measure by the presence/absence of aligning levels of work organization, i.e., the experiment, the lab, and the social world of the data repository and broader professional milieu, to ensure thresholds for data quality are met. Unless we examine the activities, practices, and attitudes in faculty-led U.S. academic research making data deposit do-able and develop theory

[12] The levels of work organization in the articulation framework were also compared with other theories of institutional levels, e.g., Scott's (2013) levels of institutions are organizational sub-system, organizational population, organizational field, societal, and world system. It is possible to map the levels to Fujimura's (1987) theory of articulation and the CMM for RDM. However, it is out of the scope of this paper to focus on comparing and determining the compatibility of the theoretical levels of institutional analysis.

to explain them, it is impossible to assess, and improve, long-term research data sustainability. Bringing these together, the three levels of articulation are the starting point and the three pillars of regulative, normative, and cultural cognitive are formed into a matrix of operational measures.

6.4 Findings

The findings document researchers' current practices and needs for research data management and data deposit. I discuss their unique, highly nuanced challenges with articulating institutional requirements for RDM and deposit. To begin, I describe the details of the participants research process, including their methods and data types, and discuss the how institutionalization shapes their research process. In this section, I describe the reasons faculty deposited data and their processes to draw out the institutional factor that facilitated and/or challenged data deposit (institutional support, learning from others, developing local, "ground-up" solutions). Then, I turn to the implications and impacts of articulation on long-term research data sustainability (researcher efficacy, data deposit success, and data quality).

I then discuss participants' struggle with reasons for optimizing their data management and deposit, such as the persisting lack of incentives and formal support for conducting research data management and deposit. These issues highlight key areas of practice for further research on how and whether policy and mandates for RDM from other fields, e.g., genomics, might – or might not – be transferrable to the social science data management context. The nature of the data, the process of qualitative analysis, and the myriad reasons why faculty deposit are part of these issues. Finally, I discuss how to better support participants' connections with their community to develop bottom-up solutions, and the need for agency over their data, despite their self-doubt or lack of institutional support for data workflow managements. From here, I conclude by suggesting an elaboration of the data articulation framework, adding a layer to the framework

— that of participant indebtedness and voice. The findings are organized according to the research questions, and illustrated by cases of data deposit, using examples and quotes from the interview transcript data and document analysis.

6.4.1 Institutional Factors Associated with Articulation Work in Research Data Management (RDM)

To address the first research question, I focus on the institutional factors that associated with articulation work during the data deposit process. Findings reveal several institutional factors associated with articulation work during the RDM and deposit process for researchers depositing data to ICPSR included those identified in the literature (i.e., disciplinary associations, journal publishers, and funding agencies). In addition to deductive coding for CMM for RDM operational measures and for "articulation" work, the analysis processes also included emergent institutional factors.

The articulation work of deposit is that which crosses between the various levels to make them align. Other factors associated with articulation work during the data deposit process included data type (e.g., Are there clear disciplinary guidelines for how data should be collected? What data types are easily standardized? What types are not? What confidentiality issues arise in data collection, storage, and sharing? What are the use restrictions of the dataset?). Here, the interviews yielded insights into ways that institutionalization facilitated RDM and deposit, grouped into three categories reported here: centralized and standardized resources for RDM, institutional support/challenges, learning from others, developing local, "ground-up" solutions.

Institutional factors intersected at levels of work organization. For example, the institutional factor of *commitment to perform* via institutional "buy-in" from the university's institutional repository to curate data requires faculty to perform articulation work to deposit data. Examples of the intersection of institutionalization of research data management and deposit in various levels of work organization are organized in **Table 17.**

Table 17: Example Measures of Institutionalization and Articulation of RDM

	Pillars of Institutionalization (Institutional Theory/CMM for RDM)		
Analytic Levels of Work Organization (Articulation Framework)	*Regulative*	*Normative*	*Cultural-Cognitive*
Experiment	E.g., Compliance with professional data analysis validity rules (CMM for RDM: Activities performed, measurement or analysis verification is performed)	E.g., Methods of data analysis documented according to open science norms for transparency (CMM for RDM: Ability to perform, resources available for documentation)	E.g., No confusion or uncertainty about how data collection should be 'properly' carried out (CMM for RDM: ability to perform, tools available to guide RDM)
Research group/lab	E.g., Taking CITI training or IRB data management certification as required by university policy (CMM for RDM: Activities performed, create procedures to train research personnel)	E.g., Newcomers to research group go through orientation training for how lab manages data (CMM for RDM: Activities performed, students required to pass a quiz after RDM training)	E.g., Shared expectations for how data should be protected or secured with lab members guide lab members' actions (CMM for RDM: Activities performed, provide data security)
Social World	E.g., Following publisher requirements by submitting data to a repository (CMM for RDM: Commitment to perform, establish stakeholder buy-in for supporting RDM)	E.g., Disciplinary association states the importance of transparency of methods in mission statement and defines outcomes (CMM for RDM: Commitment to perform, develop and justify objectives for RDM, develop communication channels)	E.g., Data repositories at multiple institutions adopt the same protocols for data deposit (CMM for RDM: Activities performed: data preservation rules agreed upon)

6.4.1.1 Centralized & Standardized Resources for Research Data Management

Formal processes, documents, and plans are institutional factors that were associated with less articulation. The reasons why these documents existed were various: it could be the institution was committed to curating data, such as at the University of Michigan (which houses

ICPSR). Less articulation occurred when labs had formal agreements and shared documents for collaborating or coordination purposes, for instance, a data use agreement, a Memorandum of Understanding (MOU), or research data management document. Since articulation work involves the work of re-aligning the goals of various stakeholders, which occurs where these are situations of uncertainty, the documents established certainty and understanding among stakeholders in advance. These clear rules and agreements helped to settle uncertainty, precluding articulation work. For example, in the case of using a core facility or a recharge center, multiple participants have reported some of the documents they use:

> *The ordering system helps you pinpoint the structure you need in your files and then [the manager] speaks to the programmer in language, the programmer will understand to set up the file structure and so it's kind of a process where like I create specifications using words, right, like I'll write a paragraph or a list, and then they take that structure and…they highlight the specific files that the information is going to come from and the specific variable names that are needed and say…exactly what to do.* (P3)

Some implications of the documents and standardized and centralized resources was making it easier to communicate with a distributed team, where the division of labor was modular. Common documents facilitated collaboration. For example, P11 explains how she created a shared document ecology to manage the moving parts across the research team and over time:

> *We have a Google Doc, a running Google Doc that we're calling our methodology report. That we just realized, when we were like okay this isn't going to be a summer project…COVID is not ending let's like formalize some of these things. And we just started adding lots of information in there about things like weighting. So we had people on the team who are postdocs and leaving and we were going to lose that knowledge. We're like 'let's start documenting,' so, for example, we had a postdoc who was preparing the data to be weighted and then working with a statistician to have it weighted. We were like 'write that down' because we're going to train somebody new on it and we don't know what you're doing.* (P11)

The PI had established expectations and workflows already. The document was an outcome from a period of previous articulation. As P11 expounds, the protocol for documentation determined "day-to-day"

> *I'm calculating the response rates and probably won't forever be calculating response rate. So I included a whole section in there about how to calculate the response rates. It's the methodology report and actually see that as something will submit to ICPSR when we do upload the data eventually so it's right now it's for us for kind of like day to day protocol, but eventually, it will be for others to see what we did, how we went about things.* (P11)

The documents referred to were often used to coordinated between centralized resources, such as core facilities, or have legal documents that describe the terms of use, e.g.: “We shared the data after the two universities signed the contract agreement, the data transfer agreement*”* (P8). Participants reported using core facilities, data analysis centers, and what some participants referred to as “re-charge” centers. These centers were used for data analysis, resources provided by the university. These documents and centralized resources are indicators of institutionalization, that is, becoming taken-for-granted and an almost invisible part of the workflow. Participants emphasized how these centralized resources indicated were become taken-for-granted, indicating a more mature RDM, in P3’s words:

> *So first off, it's important to understand that here at the [University of X] we said we have a Center and it's housed in the school of pharmacy and the center is called Pharmaceutical Research Computing [PRC]. Pharmaceutical Research Computing has a data use agreement for Medicare, a 5% random sample of Medicare data – and that's very common, a 5% sample is a common sample that CMS sells, through their contractor – so the 5% sample is housed over in PRC and they have it for years 2006 through 2018 at this moment. Medicare data tends to lag two to three years you can't always get your hands on it right away. What PRC has done – and I've been working with PRC for several years – so PRC they’re considered to be a ‘recharge center,’ meaning that they can't technically make money, but what they can do is they can't profit but they earn their money through people like me who get grants and then pay them to do stuff.* (P3)

The taken-for-granted nature of the data processes are evident in a few aspects of P3’s comment – first, that there is a center for research computing specific to the pharmacy school that has been well-established, and is familiar with Medicare data, another centralized data source. Second, the reference to the “very common” standard of a 5% sample indicates that it’s a taken-for-granted standard, one which the community have converged upon. The industry standard for sampling is understood along the division of labor, to the extent that it is codified in procedures, documents, and the technologies for data processing used by the research computing center. In addition to these taken-for-granted centralized processes, standards, and data handling understandings is a rising trend of using contractors and other means of outsourcing data activities, e.g., for survey data collection. This is a parallel with the contracting of sequencing in genomics, i.e., sending

out data for analysis to an external or internal entity. They involved the use of standard procedures, documents (e.g., spreadsheet templates detailing what was needed for the analysis), shared and/or controlled vocabulary. For example, P4, P7, P11, P12 and P14 all reported using contractors for transcription, data processing, or survey research. As P12 says:

> *[We] often use contractor, so we pay somebody else to collect the data for us, then they deliver a data set to us that has been de identified so we don't have information usually about say if it was a telephone survey we don't have the telephone numbers, if it's an survey that is an address space sample we don't have the addresses for, for instance um we may have geographic identifiers, so of course we need in terms of like census tracts or you know some kind of county or something like that, so we have to sort of keep those keep close eyes on those everything gets stored on password protected servers only people who have been approved and gone through the IRB training and all of that can get access to those data then, are kept on the servers for the lifetime of the project and then archive as needed for purposes of replication depending on what the project is and the journal, and the grant funding requirements and whatnot so that's a high level overview I don't know what you what you're interested in, but that was that's the High Level overview of keeping data.* (P12)

Centralized resources for RDM include funding, report templates, personnel and a budget for RDM. For example, faculty participating in an ethical software research project were provided with tutorials, handbooks, and resources for how to prepare data for deposit. Faculty participating also were given a stipend for the time they spent preparing data.

6.4.1.2 Institutional Pressures: Articulation Work by Faculty to Deposit Data

Articulation work appeared in situations where institutional pressures required extra work by researchers. This included formatting local data so it fit repository standards, as well as communicating with funders about the specifics of data sharing where it was not clear how sensitive data should be handled. In many instances, it was unprecedented to have an external entity mandate or even recommend a standard operating procedure for aspects of data management. Participants who deposited to ICPSR, then, had to do work to meet the demands of institutional pressures, e.g., repository requirements, funders, and journal publisher requirements for data. For example, data repositories demanded data in specific formats, in which the data were not natively produced. As participant P11 reported:

> *[ICPSR] had a lot of requirements for formatting they had a lot of specific things about you know, there was a lot of questions around whether our data would be public or restricted and what variables constituted like identifiable information, you know it's not just name and address it's like, "Well, this person is 100 years old and they live in this Community area and they're female so you can triangulate figure out who they are," so we had a lot of conversations about. That kind of stuff and how people would access the data and what the process would be, and then there was, I remember a lot of paperwork. So the process itself, I feel like took I don't know, maybe six months, and you know we didn't have like a full time dedicated person working on it either, but it was a lot of back and forth with ICPSR that I wasn't completely prepared for.* (P11)

The faculty describe shuttling between the repository, with its requirements for formatting the data, and the ways the data existed in its current form. The pressure from the repository required faculty to speak with the repository personnel then to return to the laboratory to reformat their data. There was a lack of personnel to do the data formatting and the labor-intensive preparation to make it suitable for deposit to ICSPR. It is often not a single pressure, but multiple institutional requirements that faculty manage, performing articulation work to deposit. For example, they also articulate to meet requirements to make the data findable and reproducible, as P5, a pre-tenure faculty who studies violence prevention and the sex trade:

> *You have to keep a really nice like syntax to show that that way people can see how you did things and recreate your study if they want to do, and hopefully get the same thing. So you have to upload your data your syntax, you have to upload your privacy certificate, you have to upload your data archiving plan you have to upload any final report that went to the funder, you have to upload your consent forms.* (P5)

Another example of articulating to manage institutional pressures following the funding requirements. But the funders may not specify the exact ways or places the data should be shared, leaving it up to the discretion of the faculty as to how to describe and where to deposit data. As P8, a tenured social epidemiologist described: "The funder was NIH and they have a requirement that you share your data, but they don't have a requirement that you share it in any particular way.*"* (P8) In other cases, faculty described how they had to negotiate with the funders about the specific aspects of data, when unexpected issues of confidentiality cropped up:

> *They have somebody who goes through it. Then they'll come back and they'll be like "I noticed your data archiving plans and you also have interviews like are you going to upload those are they de-identified?" For instance, with the gun study, …our funder has to approve of that so I was like let me call my funder let me talk to them get them to write a letter saying what she did has it is sufficient.* (P5)

As P5 showed here, sometimes RDM involves making adjustments to the terms of agreement and not following the funders demands to the tee and dealing with the competing rules and demands of funders, the repository, and the issues with sharing data from vulnerable populations (e.g., incarcerated individuals). In many cases, faculty feel the need to circumvent institutional requirements to protect their participants. Funding agencies and publishers required faculty to share scholarly products, such as datasets collected from vulnerable populations (e.g., prisoners, sex-trafficked individuals, prostitutes, drag artists). Sharing sensitive data is not problematic if the proper protections are put in place to protect participant confidentiality.

> *Gender ended up being a thing because I interviewed social workers and most social workers are like 80% female in the United States, but there are male social workers. And so I actually anonymized gender around some things too because I just thought "it's so rare that there's actually a male social worker," and this be somehow identifiable.* (P1)

However, faculty were not comfortable sharing research data for reasons beyond simply the protection of participant privacy. They were hesitant to share the data because of issues of decontextualization (e.g., through stripping the data of important contextual information through de-identification), breaching community trust when anyone could see and use data that was collected only after highly interpersonal trust and relationships were established with researchers or "participant indebtedness" (P5), issues of what it means to fully understand what "consent" means in sharing participant data, and "participant voice" (P1). Decontextualizing the data was an issue, especially in qualitative research. As P11 describes:

> *You also have like a very unique understanding of the data when you're collecting it versus like when you get a secondary data set that you're not that familiar with. So I do think there's some opportunity for kind of sharing across but there it's just like a completely different experience and your I feel much more like knowledgeable about the COVID data because I've been involved in every single step.* (P11)

In summary, participants who did not have a budget, personnel, or guidance on how to reconcile their local data formats with the requirements of repository metadata required faculty to perform articulation work. Faculty aligned the "messy" process of data collection, e.g., interview data,

focus groups, images, with the repository goals of creating data that is sharable, e.g., made FAIR. To address the lack of support, participants spoke to how they would create artifacts and to order the process, from data collection to data analysis.

6.4.1.3 Institutional Absence: Articulation Work and Ground-up Solutions to Deposit Data

On the other hand, articulation work also appears where there were institutional absences or an "institutional vacuum" – no precedent, guidelines, or standard operating procedures (SOPs) existed for faculty to follow, and few examples for how to manage data. This led to articulation work to algin goals of the repository and the local researchers. Participants created several documents, codifying the ways that had worked in the past to manage data, and facilitate deposit, including lab handbooks, interview study checklists, data management "orientation" PPTs for new students, medical acronym "cheat sheets," and "quizzes" to train interviewers (**Table 18).**

Table 18: Examples of the documents faculty gathered

Resource	Description	Participants	In-house/ external
Lab handbook	A compendium containing information pertaining to research group expectations for data use, file location information, and guidelines on how to store, find, or use data. Handbooks are usually stored as a digital file, but some faculty print the document (e.g., P2, P5, P13, P15).	P2, P3, P5, P7, P10, P13, P14 P15	Developed in-house
Data management plan	A written description of the data artifacts and processes expected to be carried out during the research project. Include plans for accessing, storing, analyzing, and sharing data. Includes specific ways that data will be preserved and disseminated at the project's end.	P1, P4, P5, P6, P8, P9, P11, P12, P15	Required by an outside entity (e.g., funder, publishing venue)
Data analysis templates	Files that serve as a guide for new research with a pre-formatted layout. Include Google sheets and Word documents, e.g., qualitative data analysis (P1, P5, P18).	P1, P2, P3, P7, P9, P10	Developed in-house, acquired from colleagues
Training materials	Educational or orientation documents and artifacts to train newcomers (e.g., data collection staff, students) on data management aspects. Include checklists, forms, "quizzes," and presentations.	P1, P2, P3, P7, P9, P12, P14, P15	Developed in-house, required by outside entity (e.g., CITI certification, NSF's RCR training)

Participants reported that they had few lab handbooks or lab notebooks (Table 18), a standard and widely used procedure for provenance in many of the natural sciences (Kanza et al., 2017; schraefel & Dix, 2009) and but other "template"-type documents were frequently used. These helped to organize lab workflows, as well as reduce the labor needed to teach newcomers data procedures, as in the case of epidemiology faculty member who specialized in qualitative data collection. For example, as P2 describes:

> *You can give students a template for the interviews or they're not going to all have the same font size and same formatting whatever else and like when you do this is it like the lDs to kind of be consistent, or it should be right, or I guess it's better for people accessing it if it is consistent.* (P2)

Her ecology of multiple training and procedural documents did not exist on their own, but were imbricated into the data collection processes, and carefully designed and coupled with other practices to achieve the desired outcome, in her case, data quality and thoroughness. The use of templates helped to ensure consistency and reliability for working with the data. As P11 describes, she created templates after learning lessons from previous studies:

> *We definitely had some templates when I got to [project 1]. We're using Excel for a lot of things that we shouldn't be. So we had some shared templates for like calculating hospitalization rates and I was like we should be doing this in SAS. So this kind of fell by the wayside. But then, for now, I feel like because my two big projects are so different…. When we're just focusing on analysis for the code study there's definitely some lessons learned that we have from the tobacco study that we, like the missing data, things like that.* (P11)

Over time, these RDM documents because naturalized or "taken-for-granted" – an indicator of institutionalization. For example, it become standard practice for templates for data analysis to be used. They not only assisted with the control of data workflows, ensuring data was collected consistently, and controlled according to faculty specifications, but had other results, some unexpected, such as scaling up the number of manageable students in a lab. As P5 explained, scaling up the project meant a complex division of labor and creating protocols for RDM:

> *We had to have a level of organization that we didn't have in the past. Originally when we wrote the proposal, we said that we were going to use paper and pencil to go into prisons and collect data and paper and pencil interviews and you know, in retrospect that's just silly it's just that's not the way you collect data, or at least*

> *it's not an efficient way to be able to collect data. We had to hire the right people the right research assistants and so on. And then train we had you know 70 different interviewers that we trained on this project to conduct these interviews, whether it was in person in the prisons. Because there's very limited space to do these interviews, or if it was over the phone, which is what we did with the re-entry interviews.* (P5)

The documents made it easier for the faculty PI to take on a larger number of students because they did not need to apply as much effort to train students in data handling.

6.4.1.4 Systems for Data Management: Anticipating Data Deposit from the Start of the Project

When faculty anticipated that they would need to eventually deposit data, they tended to create systems from the start of the project for data management. They were more cognizant that the data would need to be described, their provenance chain documented, and the scripts well-commented for future interpretation. This anticipatory attitude, and the consequent systems reflecting the orientation to the data's eventual reuse. In in this way, we could say research data management was not just something subordinate to the research process and practices, but something shaping the faculty and researcher activities in a fundamental way. As participants reported, the first data deposit instance led to changes in the workflows, if future data deposit was planned. As P3, a tenured faculty:

> *I think it's good if you know going into a study like "I am creating data to be shared," that may change your approach you may make sure that you're like being extremely methodological that there's nothing that you do that isn't like written down and well documented so that later there is this very well documented approach, and you know, like, I don't know if you've ever gone the National Center for Health Statistics website….all of those surveys were are always intended to be shared right it's public use data it's awesome and it's a fantastic resource, but like when you look through their documentation, I mean it is just incredibly detailed* (P3)

As such, RDM was not just something "tacked on" to the research process, but a constituent actively shaping research. Researchers re-used the processes, templates, and other artifacts they created. Moreover, they began to change their workflows when they anticipated data deposit. For example, P4 and P5 engaged in a large-scale study over 3 years with a sensitive population. The project led to an overhaul of the ways they dealt with data throughout the entirety of the research lifecycle, implementing new software tools and creating process for training new students. Their

case is exemplary of how data management can profoundly affect research, both technologically and organizationally. The nature of the data impacted almost all stages of the research data lifecycle, from data collection to storage, analysis, sharing, and publishing. The changes they initially made required substantial articulation work – there was no precedent for storing prisoner information, secondary administrative records the primary data (e.g., interview, focus groups data). For example, P5 used a dental software for their project team because they needed a solution for data storage that was not cloud-based:

> *I mean we're trying to find former prisoners in the field, and this is a population that doesn't necessarily want to be contacted. So we had to switch to a different contact management system or a CRM that was called act and it's used by you know, dentists offices and others to manage customer relations, but this is 2017 or so and everything else was moving cloud based and there was there's just a handful of non-cloud based solutions that you could keep on your server and that's what we wanted, and so we found this system called Act, which was a godsend. I've used it now for two more projects since then. It's just a fantastic system that's really customizable so we moved to that.* (P5)

P5 and colleagues had to appropriate a different system to manage data in the way they wanted, using a highly customizable system to track prisoners. They had to find public information on the inmates, data which the prison did not even know where they were. In this, there was no protocol or policy for how to organize data in such a research context. Additionally, the data that got managed was the data that *needed* to be managed, in the sense that it was seen as valuable – in other words, there was buy-in from individuals on the team, who recognized the data value. In the case of P4, P5, and P6, these were students whose dissertations and master's thesis depended on the data. The students were scrupulous about preserving the data:

> *One of the other stories that's worth telling is the management of the data as well. One of the PhD students spent considerable time working to get the right data management software that would allow us – in a relatively easy way --to begin the analysis…. I think the role of the PhD student in the process illustrates the role that [PhD student 1], [PhD student 2], and [PhD student 3] – and there been have been seven dissertations that came out of the project – each of those students (at least the first six of them) are involved in data collection data management… They really had a stake in [the data]. Their future, their career in a sense, was pegged to the success of this project. So they weren't just punching a time clock and I think that's one of the sort of unanticipated things that's resulted that we didn't plan for but we're fortunate to have had happened.* (P6)

PhD students had a vested interest in keeping the data managed. Their role was to clean and make sure the data was described and useful. In other words, before a well-established system for data management was established in the labs of participants, the data got managed if/when there was a need. Otherwise, it was taken by students "never to be seen again" (P2), or "lost forever and eaten by the file cabinet" (P8). In the words of P "too often my research data disappears into the… *nothingness*" (P8). At the publishing stage of the project, the P4 and P5 published their results and methodology in journals – the data was unique enough to merit multiple articles describing how they managed data securely, given the unprecedented and highly valuable nature of the data. After this initial data deposit and establishing a system, PIs (e.g., P5 and P6) developed a standardized system. They used the same technology for data deposit and management which they had established for the large-scale project.

> *We started at the beginning with the idea, towards the end like we knew everything that was going to go into it. But at the same time we didn't know everything that was going to go into it, because we still had to invest all that extra time, energy and money to be able to get them our archive.* (P7)

P7 anticipated what needs to be done after the project team ran the first cohort. As well, they found it a lot more efficient when you have the end in mind:

> *We started from the initial code book, all the way through the iterative cleaning and coding, it was with the idea of the end product in mind. I learned a lot from that with the idea that these next two projects that I'm working on both are they're going to be getting a lot more efficient than they were for the [earlier] project because when you have all these patterns and there's all these intricacies with the data and potentially identifying information you can't fully anticipate everything that goes into the production of a data set for public consumption. To be more efficient…we're using a different software program to collect the data… newer software programs are able to draw connections across waves better than to be able to generate skip patterns so you're being more efficient, with the way that you're asking questions, how it produces variables. So that's part of it, but also cleaning sooner, with the data, testing out your scale sooner, rather than back-ending it. That was something that we're doing better this time around.* (P7)

There are two levels of institutionalization, per the CMM for RDM model, in P7's description. The 2nd level ("Managed" process) and 3rd level ("Defined"). Evidence of a "Managed" process

within the lab (i.e., the 2nd level of CMM for RDM) is vividly illustrated in P7's comment "a lot more efficient than they were for the [earlier] project." According to P7, he began to manage the process by introducing activities, artifacts, and processes to streamline and coordinate data practices with research group during the data deposit process. However, the second level of institutionalization is less clear. Evidence of a "Defined" process at the organizational/ community level (i.e., the 3rd level of CMM for RDM) is gestured at by P7 in his discussion "…everything that goes into the production of a data set for public consumption." Here, P7 gestures at a broader community of data consumers. In doing so, i.e., in anticipation of the needs of data consumers, he responds by organizing work within the lab to meet some criteria (not defined here, but again, gestured to in the quote). In summary, the institutionalization of data management in the case of P7, here, is more clearly a case of 2nd level data maturity ("Managed" process). Yet, the community/ organizational influence is apparent in his anticipation of preparing data for public consumption.

Conversely, multiple participants already had a system in place; instead of the need to deposit data and manage it well impacting how they worked, their work was reflected in the repository. They already have system in place from previous iterations of data work (e.g., P2, P6, P7, P8). As P8 describes:

> *Because we were doing it before we submitted the archive data, the data for archiving. What we had it there, we had our own data and analysis, as I said, the cleaning the management all of that, before we even started the process of submitting the data to ICPSR so we've been managing the data for years before that.* (P8)

Yet, even in already having a system in place, there are contingencies and unexpected aspects that led to articulation work to attempt to institutionalize the data processes. For example, an unexpected part of the data collection was helping participants in dangerous situations. The work of those in social work were an example of this, where their participants and interviewers were in

precarious situations. There was a lack of procedures on how to manage the data collection situation. In the words of P15:

> *Like we did this project, where we interviewed women who use drugs and You know I say it came up more on that project and it did on like some of my other projects that are more extensively based in that kind of thing. Because women talked a lot about violence that was perpetrated against them, and that was you know could be upsetting and they talked about sex work and living on the streets and like you know there's a lot more that in those interviews that you know could raise alarm or make someone you know feel uncomfortable I mean and part of that is like Who your interviews are and training them and preparing them. And so, and a lot of times [we debrief] because the [participant] got very emotional and cried out of you know, for whatever. But so we just like talk about those things like we and we normalize talking about you know what happened in the interview. So it's not always like a big thing, sometimes it is like this woman told me she had a knife in her bag and I didn't really know what to do is like a bigger thing and we probably need to like them prepare for you know, think about that.* (P15)

The institutional factors associated with articulation work reported by the participants appear to be linked to both the transitional stage the social sciences data is in but also the places where there is a lack of institutional development for RDM. While this articulation work is notable in and of themselves, they also have important impacts on long-term research data sustainability.

6.4.2 Impacts of Articulation on Long-term Research Data Sustainability

To address the second research question, I focus on the impacts of articulation work (discussed in the above section) on *long-term data sustainability*. Long-term sustainability refers to the infrastructures necessary to preserve data and continually add value over time (Sands, 2017). In this study, participants' account of their articulation work indicated there are many important implications for several aspects of long-term research data sustainability.

While it is out of the scope of this project to enumerate a comprehensive list of the impacts on sustainability, the highlights are reported here identify areas which can pursued by future work. These highlights are: 1) *Data precarity*: data loss and failed data deposit; 2) *Data quality*: meeting criteria for integrity, accuracy, and privacy; and 3) Data sharing: Targeted sharing to communities beyond the ICPSR repository.

6.4.2.1 Data Precarity: Data Loss and Failed Data Deposit

Articulation work impacts whether and what data gets into a database at all. The ways that the dataset gets to the database is through the articulation work to meet institutional requirements, or, as discussed in section 6.4.1.3 Institutional Absence: Articulation Work and Ground-up Solutions to Deposit Data), accounting for the lack of institutional guidance for RDM and deposit. The articulation work done often is in situations where there is a lack of guidance and/or institutional support for RDM and deposit.

While aligning levels of work organization (i.e., articulation work) can *prevent* data loss, non-standardized routines for RDM become a source of precarity for data. Where there is articulation, there is lack of documentation. There can also be ad hoc approaches to describing, storing, maintaining, and, ultimately, sharing data. This lack of documentation and ad hoc approaches were apparent in participants reports. Participants reported their experiences with data loss, and the ways that articulation work could attempt to mitigate it. For example, P8, a social epidemiologist (tenured) described data loss:

> *I've seen too many data sets in my academic life that disappear into the…* nothingness*…and I don't understand it. I feel researchers never do all the papers they think they're going to do. It was always clear to me that we had to put in the data [to the repository] … and because I thought it was a very special data set. I mean we had incredible response on that…deposit. Almost 15,000 [downloads]… There was like huge response to both of [the datasets submitted to ICSPR]. I knew that I expected that would be the case and that was the case* (P8)

The lab members had to articulate the alignment of multiple levels of work organization to make deposit possible: locating the data at the experiment level, identifying the repository requirements at the social world level, and coordinating with their research group through at the lab level. Similarly, the work of preventing data loss is through articulation – without the work of making sure data is depositing data do “disappear into the nothingness.” Data loss and failure to

deposit was reported for 8 of the 15 participants (P1, P5, P7, P6, P8, P10, P13, and P15). As described by P6, a professor in criminology:

> *We have we almost lost the data – one of our PhD students who knew the data up down inside and out got a job and that's great we want them to all get jobs, on the one hand, on the other hand, getting those data put up at Michigan [at ICPSR] was something that she continued to help with despite the you know the grant that supported her largely been over and we had to come out of our own pocket for all of that yeah there was no funding to support that and that was that was expensive yeah and I think you know for the process of archiving funding agencies may want to give consideration to having to set aside. In addition to what it costs to do the research but and my final point about the archiving is one of the virtues of the delay was it allowed us to get these five methodological articles. Out there that kind of establish so now young researcher and older searcher who wants to pick those data up and do some analysis can say, based on the […] reliability and validity checks, we know these data meet those standards and we think that's also a service that that comes with the data.* (P6)

In this case, P6 reported almost losing the data because there is not institutional continuity, e.g., through a lab manager. The reasons for losing data and failure to deposit were wide-ranging, from lack of personnel and budget (P1, P3, P14) to a perception that the data was not valuable outside of the immediate research group (P2, P9, P10). This work was highly specific to the characteristic of the qualitative data, such as the confidentiality requirements were higher and training needs are elevated. P8, a tenured professor of social epidemiology at an ivy league institution, continues:

> *We don't have the resources to even begin this process [of data deposit]. Mostly personnel because you need people to be able to read it, you have to train them, and we haven't even determined what we would need to do to determine confidentiality. Because in qualitative interviews it's not that straightforward, I mean doesn't have names and identities, but like I said, you have to look through it. If it's in a large city, there might be different issues, then, if it's a small location in terms of how identifiable a person might become* (P8)

Several participants cited the extra work it takes to prepare data for deposit. For example, it takes a lot of work to identify the requirements of repository, then to clean, format, and describe the data. Cleaning and merging the variables can take days and cleaning data takes a substantial amount of work. As P7, a tenured faculty of criminology who works with administrative, focus group, and interview data reported:

> *You've got to build in a half a day just to get all the right variables to merge because sometimes the data sets get changed. [On] our recidivism data set we'd do extractions from the Department of Public Safety about every six months and all of a sudden, somebody might have used different line of code, or a different title and it could get pretty messy to be able to merge them. Or with the official datasets just something silly like the prison would change a variable name and then you have to go back through and try to figure out what the problem was. I mean it was it's a big headache. It's like a half a day you gotta devote every single time you want to start up a new project* (P7)

P7's experience shows that the burden of data preparation for deposit is high, even though it may be valuable to standardize the data, e.g., by reconciling longitudinal data. Without the personnel or budget support, given this high bar of work, many datasets do not get deposited, and fall into the "long tail of dark data" (Heidorn, 2008). The impacts of articulation – and the lack of routinization – can also be seen in the issue of data quality, discussed in the next section.

6.4.2.2 Data Quality: Meeting Criteria for Data Integrity, Accuracy, and Privacy

Implications of articulation on long-term sustainability also includes data quality. Data quality encompasses not only the accuracy of data, but also additional criteria such as preventing material decomposition and data deprecation – i.e., *data integrity* (Herzog et al., 2007). In addition to this broader scoped definition of data quality are the activities taken to add value to data, such as making data interoperable by adding metadata to describe datasets or normalizing the variables, e.g., putting quantitative values into a compatible scale or translating qualitative data transcripts into multiple languages. In addition, data privacy and maintaining confidentiality for participants added value. Meeting repository criteria is mediated through ethical, material, and epistemic thresholds for research data.

Participants ensure data meet the criteria for data quality (e.g., integrity, accuracy, security) through their data management practices using an assemblage of artifacts (e.g., documents, code, data), equipment, and labor. The implications of articulation work for data quality were made clear in participants' accounts of working with their research group, e.g.,

students and collaborators. Some participants contracted to ensure the quality of their data met these checkpoints. P13 described:

> *I did not do all the checks, I had caught, so we are actual survey administrator we had contracted with a national research Center. They run a panel called* America Speaks*, who operates and utilize them on other studies as well. When I got the day initial data from them, we you know to our benefit inherited many of their protocols and processes for our project, which are guided by best practices and a few other kinds of, not accrediting bodies, but kind of professional oversight bodies. And so, all of that is in the documentation as well around as well as includes some of the details of how their panel is structured should anyone, you know really want to get into those details of how they create their lists and weights.* (P13)

As with P13, participant P15 used external standards and guidelines as part of her checking for data quality process. Her process also reflected how she "inherited many of their protocols and process," but instead of from a professional oversight body, it was from her training as a clinician:

> *Like for me, like always an evolution and then you have to keep in mind, obviously, like, whatever your institutional standards are, but I think institutional standards are the bare minimum, for the most part and there's like a lot more that you can do to make your data more robust more rigorous more better documented, I will also say, the thing that probably influenced me as I came from the clinical trials world where like documentation is you have auditors coming in and auditing like medical charts and study documentation, and so I came from a place or I came from a background where documentation was like so take tightly monitored and so that probably influences like how I think about documentation a little bit as well.* (P15)

Likewise, P6 reported that it was a "godsend" that they were able to draw from the work of a previous project that used similar software, and could inform their data management process: "There was a large study in in in criminology that was done before us that used it and that study started in 2000 and continued to like 2007, and so we drew on their experiences from blessed" (P6) Often, as with preventing data precarity and institutionalizing RDM (see 2.1 Institutional Factors Associated with Articulation Work in Research Data Management), data quality was facilitated through standardized artifacts such as shared documents, e.g., lab "handbooks." For example, lab notebook documentation was a way faculty made sure the quality of data was maintained through the research process, from data collection, and analysis to file copying, storage, security, and de-identification. As P10, an assistant professor whose research focuses on health services barrier and substance abuse described:

> *There are a number of kind of guides that we have for that we used when the when the lab was solely focused on that project and then that we use now as people come into work on it right, so we have basically a handbook and PowerPoint forum that is an overview of the entire project…. I also do have a document that sort of authorship our guidelines…And this is a way to kind of make sure that you know people are stepping on each other's toes for particular you know topic or questions. Particularly because we have a lot of data and so people if people work on it for years, making sure that No one has addressed this before, who are the other collaborators, going to be who might want to be brought into this. There's sort of a process to that, although I will say that I that's a little more body in terms of its use, I think it depends that was being used a lot more also when we had projects …that multiple faculties and students were working on.* (P10)

Moreover, the intertwined nature of data quality as *accuracy* and data quality as adhering to agreed-upon standards and values for participant privacy is demonstrated in interview data on RDM workflows. For example, P2, a pre-tenured sociology faculty member noted how the nature of data collection led to precarity, not only of data loss, but of provenance maintenance and issues of data security. He comments on the extent to which his data management processes are systematized:

> *It was really bad… [For] my dissertation I was traveling around Iowa doing interviews and I was using my phone as the recorder and I would go from one [interview] to the next to so and I'd be driving across the state to go next so there wasn't really like that ability to kind of like "okay, I just did my interview now let me like put this on a USB." Obviously, there's certain protocols you do have to follow in terms of security… But my consent forms, for example, if I'm driving around, they're my car with me, right?* (P2)

The data P2 maintained was done through ad hoc file keeping, but also meeting the requirements set forth by the community (i.e., as reflected in the IRB). The privacy of data was enabled through the steps he took to backup files – preventing data loss – while also managing the privacy of his participants. The articulation work done to use the physical space of his car to manage research data impacted data sustainability. Articulating between the various data regulations, uncertainty of traveling to collect data, and intertwined needs for data backup, security, and participant privacy had important implications for data management. They implicate the value of data for sharing – for the immediate community, as well as communities extending beyond the ICPSR repository, as discussed in the next section.

6.4.2.3 Data Sharing: Making Data Accessible and Sharing to Broader Communities

An additional implication of articulation work in RDM on long-term sustainability is in the realm of *data sharing*. Data sharing, especially sharing through a repository, is an essential component of long-term preservation. Preservation relies on a centralized location to access data, as well as governance principles that make the data FAIR (Wilkinson et al., 2016). The participants in this study reported how they shared data to a repository, surfacing the articulation work of tinkering and doing extra work to meet the requirements of multiple stakeholders. For example, to share data, participants had to meet the requirements of the repository, which were sometimes difficult when the deposit technology isn't "user-friendly," as described by P2:

> *It was a little confusing – I was going to create a DOI for that, but I don't find ICPSR like the most user-friendly easy to navigate. it's a data base with datasets yet, but it's I think it's the best one out there in terms of like having a lot of data sets being available to access.* (P2)

However, routinized processes (rather than articulation) often assisted in data sharing, such as the IRB requirements for sharing. Yet, even routinized process required articulation work. For instance, as P1, a bioethics faculty working at a large medical institution described:

> *The amount of stuff you need to do from IRB standpoint for doing your own institution with their institution and if I wasn't already at [university] with ICPSR, so a lot of instruction around making sure that we were reading those regulatory pieces.* (P1)

Learning the "regulatory pieces," such as P1 described with the IRB, was important for making data deposit possible, impacting long-term sustainability. The articulation work of sharing data positively impacted data sustainability when it came to sharing data beyond the ICPSR community. For example, P2 shared the data on their website because they knew ICSPR was not accessible to many communities. That is, the datasets were not accessible to some people without an institutional affiliation. In addition, they may not have appeared on popular search engines (e.g., Google). As P2 elaborated:

> *In terms of people finding [data] that are not like looking for data sets and searching for data sets like and coming across it or like somebody who's just interested in drag or interested in a particular draggers might want to read or something like they're not going to necessarily be looking at ICPSR. They might Google Tomahawk Martini and then they can like read about them here or they might be doing a search for a lit review for some type of project. But like I feel like in most cases, unless they're looking for a dataset they're not going to come to ICPSR. They're probably also not expecting like a drag artist interview data set on there, anyway. What [data] I posted on [professional website] I mean that'll show up when you do a Google search whereas I don't think this [ICPSR data record] would.* (P2)

Faculty are dealing with governance uncertainties through this "articulation work." Amidst the shifting landscapes of legitimacy of datasets, and the still-forming legal rules and data governance. P2 recognizes there are rules for providing access to journal publications – often there are access control restrictions. With datasets, the "publishing" analogy for datasets does not hold because of legal, accessibility, and sharing differences (Borgman, 2007).

> *I think this is great like this is kind of like a central repository for a bunch of stuff right I know more, not necessarily spark because that's isn't official repository but like just like putting on my website, this is like more legitimate in some ways right but I think that it's great to have it in more than one place, as long as you're legally allowed to do so, which I can with this right, but like you know, like that's the issue with journal articles and stuff right, I was like where do people posted and what kind of access, can you have or not have But with this right, I can I can put it both places and that's great because then it hits different people depending on… how they kind of access, material and look for things.* (P2)

The central repository was a key part of data sharing. Yet, the nuanced understanding of faculty as to how datasets are regulated led to faculty articulation work to make data accessible to a larger population. The work P2 did to post data on their website reflects an understanding of the technological, regulatory, and social behavior of the potential data reuse population. The technological understanding is that search engines like Google only display certain results for user – and ICSPR data may not be in the top-ranked items. The regulatory understanding is that the data may not legally be able to be posted in multiple locations. He also anticipated the information behavior of some users – namely, that they would most likely access the data not using ICPSR. The articulation work is both material – posting the data to his website – but also intellectual, in his consideration of the multiple components of making data sustainable over the long-term, in this case, by making data accessible to wider audiences. The institutionalization of

articulation, and its impacts on such examples of sustainability are further elaborated in the next section which discuss the findings.

6.5 Discussion

The findings reveal the complex feelings participants had toward the practices involved in the institutionalization of data management. On the one hand, participants saw the value of institutionalized data sharing and even created their own rules and policies for data sharing in their research groups, such as templates for data analysis. On the other hand, they also liked the freedom of the process, the ability to control how data was collected and shared, resisting organizational changes toward increasing institutionalization and standardization. This tension between the desire for institutional support for RDM assistance and the need for agency over data management suggests opportunities for greater connection between policymakers and researchers to develop qualitative research data sharing.

The tension also suggests that the regulations put in place always require work to be carried out in practice. In practice, participants did always not conform to institutional mandates, nor did they "optimize" their data management processes or always apply best practices to their data management (Qin & D'Ignazio, 2010). Furthermore, no matter how detailed the requirements of institutional guidelines, they must always be customized to the particularities of local practices. To unpack these tensions, I turn to a framework proposed Gerson and Star (1986) to describe the due process in information systems by making "local adjustments that made the work possible in practice" (Gerson & Star, 1986). This framework helped to unpack how *institutionalization* (W. R. Scott, 2013) and *articulation work* (Fujimura, 1987) intersect.

6.5.1 Due Process in Information Systems

With the rapid institutionalization of data sharing, how are researchers reorganizing their data workflows to accommodate it? Scholars have theorized the work people do to locally in response to institutionalization of a task or process is what Elihu Gerson and Leigh Star call "due process" (1986). Due process is defined as “articulating alternative solutions.” (Gerson & Star, 1986). A major part of due process is reorganizing workflows to accommodate the abstract to local circumstances, also known as a form of articulation work. When some aspect of life shows indicators of institutionalization, there is a level of work that must occur to implement the theory to the practice. Institutional theorists call this institutional work.

Articulation work is one facet of institutional work. In exploration of this issue, this study examines how scientists do ‘articulation work’ in response to pressures from institutions to manage their data such that it is ready for deposit and how they reorganize workflows to deposit data. The tensions between highly specific local practice and decontextualization as the value-add proposition is vividly illustrated in social science data management and deposit. That is, data which is sharable because it has been “de-contextualized” (Ankeny & Leonelli, 2015) and made “pristine” (Plantin, 2019) for deposit. For example, participants valued the flexibility of managing data according to the study questions and goals, as reported by P2:

> *I tend not to have a set framework, so it really is based around the study and the questions for kind of going through the different transcripts and seeing overseeing. And then, like teaching them to hold for their existing knowledge within emerging knowledge it's happening and that tension and being reflective there. And then often they've never written something up either so taking that data and synthesizing it and then putting in a report. P2*

Articulating data institutionalization is a concept that is developed here, to refer to the processes whereby data practices or values become taken-for-granted and legitimized through the work of customizing underspecified implementation conditions by making those “local adjustments that made the work possible in practice” (Gerson & Star, 1996) to align levels of work

organization. Articulating data institutionalization activities include “considering, collecting, coordinating, and integrating” (Fujimura, 1986) and “scheduling subtasks, recovering from errors, and assembling.”

6.5.2 Implications for Policy and Practice

The findings also underscore the need to design appropriable tools that can work with unanticipated workflows and allow for the flexibility to work with existing local practices in a “communicative ecology” (Gonzales et al., 2015). This study surfaces a complex interaction of where the institutional features of the research environment intersect and mesh with the work of managing and depositing data to open research repositories. Initiatives for depositing data to repositories: the case of ICSPR were associated with high levels of data deposit. For instance, participants’ data deposit practices were often associated with formalized programs such as the Qualitative Data Sharing (QDS) Project Series and the Washington University (WUSTL) project which paid participants to submit their data and improve an AI-enabled software for de-identifying and anonymizing data. The presence of funding requirements was highly associated with data deposit. The implications for policy are that these discipline-specific initiatives can spur greater data deposit, but also involve scientists at the level of data policy and system-design. Such research buy-in can help to promote their voluntary participation in improving long-term data sustainability efforts (as we see in the ground-up activities of the genomics community to establish model organism databases).

There were tensions between the goals of making data FAIR and protecting participant privacy. The participants were ambivalent about this as well; there is a still-forming discussion and debate about how to reconcile open science goals of democratizing data while also

protecting the vulnerable populations' data. P5 described this tension in her work with sex trafficking victims, and serves as an exemplary case of this ambivalence:

> [The data] is very detailed vulnerable stuff. It feels to me like this person that I interviewed trusted me… *I formed a relationship, I did the due diligence of spending time in the community, to get the community's trust… to gain the trust of this person to have her allow me to do an interview about her abusive childhood and sex trafficking experiences. I don't think it would be fair [for] some random person that she doesn't know who lives in like, Chicago, being able to access that and write their own thing about this like feels not ethical to me. I go back and forth because I'm like "Yeah data should be like people should have access to it" …but not all data. Because …I don't know what that person is going to do with this woman's story. And for them to write like a journal article about this for their own career when they didn't have any actual interaction with the person and didn't do the due diligence of building trust and making promises about how their narratives we're going to be used...They made themselves vulnerable to tell you I feel a little uncomfortable with.* (P5)

Standardizing the data requires decontextualization (Leonelli, 2014a), which necessarily depersonalizes the story and narrative of a study subject, as P5 describes. These, and the work of due process, need greater empirical examination and community discussion, to form sustainable solutions for making data FAIR as well as CARE – an alternative framework that prioritizes Collective benefit, Authority to control, Responsibility, and Ethics (CARE) (Carroll et al., 2021). The CARE framework is one such approach to addressing these tensions, and complementing the needs of both participants and research communities.

6.6 Limitations and Future Work

This study focused on faculty in U.S. R1 institutions who deposited data successfully. While participants reported instances where they failed to deposit data, the population was, in large part, successful in their efforts to deposit data. One limitation, then, is it does not identify the major barriers when data does not get deposited. As well, the findings suggest that not only is research data management a product of the research design, but also vice versa. This reflects existing work (e.g., such as STS and CSCW) showing the reverse can be true: that data management practices can influence research design and even topic selection (Bishop & Hank, 2018; H. Collins & Pinch, 2002; Latour, 1987). As mentioned by However, this study did not

take a longitudinal perspective to show how institutional logics, policy, and technologies influenced the ways that RDM influenced workflows. This is a limitation which can potentially be addressed in future work, for instance, focusing on co-constitution on research data management workflows, technologies, and research design.

Another limitation is the choice of subject sample. The subject sample focused on faculty who deposit to ICPSR. The ICPSR repository is institutionalized. At first, the ICSPR may appear to reflect a highly mature institutionalization of data deposit. However, within the research group, it was clear from the data analysis and findings of this study that the presence of the institution did not reflect a universal uptake and practice mature, highly managed data practices. For example, even among scientists who successfully deposited data, the research group did not demonstrate shared metadata schema, common rules and norms for how data should be handled, or in other words, the processes social scientists employ within their lab are not highly institutionalized even though ICPSR has created a mature infrastructure for data deposit. Thus, another limitation of the study is that the participants do not face pressure to submit to data repositories.

Prior studies show that data deposit and sharing is much more fragmented and scattered in the social sciences (Jeng et al., 2017; Yoon et al., 2016). These studies often use ICSPR as an example. So, the choice to focus on ICSPR is justified by prior literature, and yet, the choice does limit the study generalizability to the ICSPR context. Other repositories have different requirement and regulatory structures, impacting the data “articulations” of scientists (e.g., Figshare, Harvard’s Dataverse) (Lai et al., 2011). However, the initial findings may potentially apply to those with similar governance structure (e.g., repositories with staff who check datasets, require metadata descriptions, are the “go-to” repositories for a field or journal)..

Future work will address how these data management process, and routines, crystallized. In this stud, we found routines were established when there was a need to coordinate as in an interdisciplinary collaboration; where standardized, shared processes facilitated data management. While the practices described by faculty here have impacts on the reliability of the infrastructures for data preservation, there are still many open questions as to *how* these routines came to be. The evidence of institutionalization of RDM can be seen across the ways in which articulation work impact data precarity, data quality, and data sharing. Yet how they normalized as routines at all is an open question. Future work could explicate how routines become institutionalized to better support research data management and inform data policy.

6.7 Conclusion

This study examined how research data management and deposit is institutionalized in the social science context. The study used interviews and data deposit records to surface the institutional factors shaping research data management (RDM) and deposit. The finding show how research are incorporating institutional policies and norms into workflows, and the implications for long-term data sustainability. I discuss the potential ways that less-institutionalized contexts can learn from the more institutionally mature disciplines to make their data findable, accessible, interoperable, and reproducible (FAIR) (Wilkinson et al., 2016). I discuss the implications of the findings for long-term data sustainability and issues of institutionalizing data management and deposit, and develop the concept of *articulating institutionalization*, drawing from the *due process in information systems* work (Gerson & Star, 1986). Recommendations for data policy include incorporating scientists' practices and processes into the design of data deposit repositories and processes. Future research directions can examine how institutional components establish and the logics that reinforce them.

With these understandings, researchers can better isolate variables and develop models to support data deposit. Science policymakers can plan interventions and allocate resources to prevent data loss to for long-term research data sustainability and professional organizations can assist with faculty development. This paper also contributes methodologically by identifying core measures of analyzing the extent of institutionalization of research data management to compare across disciplines. More broadly, examining the work of data deposit disabuses us from the grand vision of cyberinfrastructures by helping understand the work involved in data management and deposit in which scientists engage to meet emerging institutional demands.

CHAPTER 7
DISCUSSION OF FINDINGS & CONCLUSIONS

This chapter discusses the main findings and articulates the research questions (RQs) answers, driven by the central RQ: *Why and how do faculty deposit data to research data repositories?* Connections are drawn between the findings of the three studies when taken together, and their implications for theory and where there are gaps, pointing to opportunities for future work. This chapter discusses how genomics and social science data management contexts compare, highlighting important differences in data practices between the populations. Then, I articulate the main contributions and the implications of the findings for stakeholders including data repository personnel, funding agencies, and policymakers. The limitations are identified to highlight the constraints and opportunities the selected research design choices enabled, and the impact of the COVID-19 pandemic on the conduct of research and interpretation of findings. The chapter concludes with opportunities for future research.

7.1 Discussion

7.1.1 How research is shaped by data management practices

Across the three studies, a major unifying thread that emerged was the relationship between the research design (e.g., research questions, methods, scope, hypotheses) and data management practices (e.g., data cleaning, storage, dissemination). Now, the conduct of research, and its design choices, are often seen as independent of data management. Data management is seen as subordinate to – and in service of – the research design, having no influence on what questions are asked, the methods, or hypotheses made. How could data management, mundane but necessary part of research, shape how researchers design their studies? Initially, it might be intuitive – the research questions, the hypotheses, drive the

collection of data, the organization of data, not vice versa. However, the findings of the three studies suggest that the opposite can occur, as well: research data management can influence how and what research is done, the questions that are asked, the scope, and the hypotheses.

Study 1, an interview study with molecular biologists in genetics and genomics showed how data deposit can limit or broaden the scope of the research study, because the repository often asks for more data (e.g., "show me the negative data" P11). We saw in study 2, also the genomics case, that the pressures to deposit lead to researchers anticipating data deposit from the start of the project. This anticipation can influence taking different actions with respect to ways that questions are asked. Anticipation work has been theorized as an important planning activity, one that has not been acknowledged as having sway on the organization of work in scientific labs. Steinhardt & Jackson (2015) theorized *anticipation work*, the work of looking ahead to envision what is needed now anticipating future goals and events. Anticipation work has links to "articulation," the theory which was applied in Study 2 in the sense of *data deposit as articulation*. The notion of deposit as articulation work was developed further in study 3 with a novel population. In study 3, the novel population – social science faculty – related the ways that depositing data introduces new types of work and implicates the everyday design practices.

Across the three studies, we saw evidence that the practices and work processes and routines of managing and depositing data to repositories shape the ways faculty plan their research, including selecting research questions, generating research hypotheses, determining the analyses which are feasible, and scoping a study. For example, P11, a tenured faculty of clinical public health and social epidemiology reported how she saw the process of data cleaning and management as a novel form of work but also one that led to generating hypotheses:

> The *data cleaning and management is the most time consuming, and people might view it as the least fun. They just want to get to the analysis. but your analysis won't be valid if your data isn't correct. And I actually really like the data cleaning and management process because that's when you learn about the variables and*

> *other questions come up and I feel like that can be hypothesis generating… When I didn't have a specific hypothesis, I would just do simple things to clean and look at the prevalence. I remember texting my coworker like "I can't believe the prevalence of smoking is 50%." We were going back and forth about all of these basic findings that we would have missed if we had been like "I really want to know if X is related to Y and I'm just going to go straight to the regression. So I think people don't think of it as necessary or interesting, but it can be vital.* (P11)

For P11, the research hypotheses were influenced by RDM. As P11 sifted through the data, she learned more about the variables RDM. Generating hypotheses is an intellectual, scholarly step in the research process. However, it remains a dominant perception in the literature on research data management that RDM is a "mindless" activity that an aside to the primary "intellectual" work of doing science. Yet cleaning the data is not a rote, "mindless," activity lacking scholarly content. Rather, the RDM has "intellectual" content, contributing to the research process. Tukey (1977) highlighted this point, showing that cleaning data – a part of "exploratory data analysis" –generates hypotheses, by doing variable inspections and managing the data (Tukey, 1977, p. 23).

What are the consequences – and implications for long-term sustainability – of how RDM influences research decisions and practices? Prior research shows databases influence collaboration dynamics in genomics and related fields. For example, Bietz and Lee (2009) posit databases impact the organization of work. Bietz & Lee (2009) show database are sites where researchers are forced into contact, leading to conversations about how to address the legitimacy of methods, and what data should be deposited. Here, they argue databases serve not as boundary object, but are better understood as "boundary negotiating artifacts" (*ibid*, p. 3). In developing the databases, scientists needed to come together to make choices about what metadata to use, the tools to use, the query and search systems, and what data should be deposited. As the authors highlight, these seemingly technical questions were opportunities for scientists to declare their epistemic commitments to the legitimacy of certain methods, of which questions mattered, and

comparing the validity of analysis methods (Bietz & Lee, 2009). Databases in metagenomics existing across communities served to organize scientific collaboration. In the case of data deposit in genomics, the RDM changed their lab practices, internally, whereas in metagenomics, the impacts were more external, with outside contributors to the database.

While there are many aspects of research data management and the research lifecycle that were implicated by the introduction of RDM, here I highlight one: anticipating data deposit from the start of a project. I highlight this one because it has potential for future research on the implications of data policy and systems design which incorporate artificial intelligence (AI) into computational research (e.g., the interpretability of neural networks and data cleaning; the ways that AI can be black boxed in software for de-identifying human subject transcripts). For example, participants reported how anticipating data deposit the lab "rode the momentum" of well-organized data (P4). They explained how anticipating data deposit led to the conduct of longitudinal studies. Here, RDM shaped their impetus to do longitudinal studies. In other words, through exposure to the opportunities databases offered to standardize, link, and securely store their data, faculty saw the potential for transforming their current, single-year studies into longitudinal studies using the databases functions for FAIR-ness (Wilkinson et al., 2016).

7.1.2 Making tacit knowledge explicit: learning to manage and deposit data

The findings across all three studies revealed researchers need to learn to manage and deposit data. It became clear that while some of the knowledge is easily transferred through reading documentation or “Googling it,” the knowledge for how to manage data and deposit data were *tacit knowledge* – that is, information which is not codified and not easily learned through indirect experience (Jennex, 2009); one cannot gain tacit knowledge, for example, of the most

effective angle to hold a scalpel in the operating room by reading from a textbook. Rather, tacit knowledge is learned in context through practice and apprenticeship-style transfer of knowledge.

The research findings from Study 2 and Study 3 especially underscored multiple points of learning necessary through the data management lifecycle, many of which involved tacit knowledge about RDM. For example, faculty spoke to how they and researchers in their lab did not have codified data management tools or documents to guide their lab's data management. Faculty had to experiment with setting up a lab with the tools that will work harmoniously across the RDM lifecycle. To acquire data, store it, ensure its security, and enable analysis and dissemination is on ongoing process, improvisational, and parts of which are explicit knowledge and parts of which are tacit knowledge. For instance, researchers described a struggle to keep up with the latest tools for data management. Students needed to learn from another lab to acquire the skills for their lab. P2, a genomics faculty (Study 1) described this as "learning by doing" by visiting a lab specializing in a specific technique in X-ray crystallography. As he explained:

> *If we need to learn something from their lab then we either somebody from the lab will go or I will go, usually not me, but a graduate student or a postdoc. They will go over there and they spend a semester, a month, a week depending on what we need to learn, and learn that technology or method and come back. They use it yet sometimes we can just send the material to them and they will analyze it, they will analyze it and send the data back. (P2)*

Their students or postdocs would be tasked with trying new technologies, e.g., for collecting data and processing it, by visiting the labs for a period. P2 would send his students to another lab to learn how software systems worked – in large part for data analysis – but in part which included cleaning and processing the data. The processual knowledge – ranging from formal to informal –learned at the other lab for dealing with the data would then be taught to others in the lab or kept as knowledge possessed only by the visiting student.

Moreover, since data are often digital – even material samples are represented in a digital format – researchers rely on technical systems to manage their data, such as servers, databases,

and repository software. Learning these tools was a common theme across the three studies. For example, in Study 2, a researcher in molecular biology mentioned how the use of Illumina was a new tool for managing and cleaning data. Implied in his response is that when the system was introduced within the sequencing workflow, he or a member of his lab learned to use the interface and its functions, given the necessity of engaging with the platform to retrieve and clean data arriving from the sequencing center (e.g., the Molecular Core Facility):

> *They [the core facility] carry out the next generation sequencing runs then they'll let us know when it's finished. And then we can log in and see a 350 million pieces of data, which is a little overwhelming…I think they're changing them, but, so, when we use something called BaseSpace which is, I think it's part of the Illumina platform, it's just kind of a nice place to keep your data organized. You can process it there. And then they have all these tools you can use for extracting information and all that stuff.* (P1)

P1 also referred to the proliferation of tutorials on how to deposit data, such as the documentation, handbooks, tutorials, and workshops provided by GenBank, through the National Center for Biotechnology Information (NCBI). As well, faculty referred to the learning processes as a technical barrier for entry, requiring researchers to learn frequently, or to stick with what they know at the chance of losing out on a new technique. Faculty are thus tasked with learning – or having their students learn – the changes in information systems (e.g., new fields added to data deposit forms, file format changes). Policy requirements also change, pressuring faculty to keep up to stay in compliance. Through the process of learning about data management and database requirements, research generate substantial knowledge about their data and phenomenon which they are interested through data management. The tacit nature of information of learning to organize and secure data is described by P5, a pre-tenured faculty in sociology:

> *In the IRB it says how we'll protect the data. So that's already set out and it's sort of like common knowledge about how we do that at our organization. But I don't think we have any kind of* written *anything like that. It's just more like depending on what the project is. It's like "Okay, you've got audio files, you upload them to this private drive that only you know your team has access to." There's no – it varies by project and there's no real specific protocol, except for what we've outlined to the IRB.* (P5)

As P5 reports, there are ostensible and performative routines embedded in RDM (Feldman & Pentland, 2003). These routines are codified in a general manner in documents as *ostensible* routines, and performed in practice by the researchers as ongoing *performative* routines (Feldman et al., 2016). The routines literature can help to illuminate the ways that research data management involves learning, codifying, and transferring knowledge. Researchers can learn tacit information from performing the tasks, as they incorporate the tool into their workflows. Depositing data, similarly, requires knowing – or learning – the tools available for data deposit, such as the databases available to submit to, and how to make their data fit the file formats and template standards required by the repositories.

Research data management (RDM) is still widely seen as maintenance work, or an aside from the "primary" work of science. It is the extra work that is taken-for-granted as faculty responsibility. As one participant said, we do lots of things we do not get paid for as faculty members. I already have an "ambiguous relationship" with academia. Data deposit is one of them that if I am going to do something with my extra time, it is not making data clean to deposit it (P2, Drag): "Unless it is prioritized and supported, it is not going to happen. More well-resourced institutions who can afford PMs [project managers] have them. The project managers do data organization and management work. As such, they possess much of the taken-for-granted information necessary for dealing with data. For example, a tenured faculty in genomics (P3, Study 1) described their lab manager (their version of the project manager) as the "lab mom," who teaches students how to input experimental results into lab handbooks and to organize data.

In work on science and memory, scholars have discussed the relationship of institutionalization and scientific knowledge. For example, Baker & Bowker (2007) discussed how the "total institution" eliminates the need for conscious memory. As (Douglas, 1986)

argued: “when everything is institutionalized, no history or other storage devices are necessary.” The institution does not need keep any records about the student or the prisoner except for registering that they exist as part of the system. For the period that they are there, there is no reason for them to remember anything about my own past because the institution ‘remembers’ “all it needs to know through the complex set of procedures that it puts into place” (K. S. Baker & Bowker, 2007, p. 128). The inscribing of tacit knowledge into the workflows of social science labs is how the group remembers. The continuity of the work depends on memory practices. As data management becomes more institutionalized, memory practices will likely shift to the “complex set of procedures” and regulations an institution. The implication for scientific work is that tacit holding of knowledge may be offloaded to such systems, implicating the traditional approaches to scientific training such as the *apprenticeship* model.

7.1.3 Disciplinary distinctions regarding data deposit expectations

There are important disciplinary distinctions between the sample populations of Study 1, Study 2, and Study 3. Genomics and the social sciences are fields characterized by different levels of cyberinfrastructure “maturity” – which encompasses technologies, policies, practices, and process for RDM. From the three study findings, several important themes emerged from the of the distinctions between genomics social sciences disciplines.

Noted that the impetus for selecting the two disciplines was to compare different contexts for data management: high-institutionalization (genomics) and lower-institutionalization (social sciences) of data management and deposit. This comparison (premised on an assumed distinction between the research environments of the two) was motivated to use genomics’ higher-level of institutional maturity to inform the social sciences to potentially inform less-mature contexts of the institutions and institutionalization of data. A driving question, here, was: What can the social

sciences learn from genomics' research data management, and vice versa? Does what is effective in one field work in another?

First, there were important differences between genomics data deposit and the social sciences. For one, the confidentiality and unstructured nature of the social science data created clear distinctions between how the data were managed and deposited. The encryption and security requirements differed, as well as the expertise needed to analyze the data. However, a similarity between the two was that anticipation of deposit resulted in more mature RDM. Comparing genomics and social sciences reveals that a major overlap was the extent to which RDM was established as part of faculty's lab if they expected to deposit the data. However, the ways that this manifested was different between disciplines. As P11, genomics faculty in plant pathology described:

> *Nowadays with the whole funding...because we do a lot of genomics, in the genomics world for publishing the genomics work, you have to submit it to a publicly available repository. If we do a lot of genomics work and if we do some gene identification sequence, that gene, it has to be submitted to the database called Medline, PubMed. That's where all the sequences have to be submitted before it can be published... when we started out the procedure was not very clearly laid out and people had their own ways to collect the data and so you have to translate it into language which is acceptable to the repository. Now people have learned it. So it doesn't take that long. Depends on what kind of data you have, how much data you have. In our lab, it is not that complicated to submit the data. There are already very standard protocols, very standard procedures for how you create your data and once you have it in that format to submit.* (P11)

The faculty who initiated the project with the end deposition in mind have more artifacts to coordinate the process. Lab handbooks, how-to guides, and analysis template are used by the students and other researchers to coordinate their efforts and standardize data descriptions. In the case of social science researchers, as a result of anticipating deposit, they changed some of their workflows. What became clear across all the studies in this dissertation was the plans researchers made to anticipate depositing to an open research repository. For example, as P3 (Study 3), a tenured faculty in criminology describes:

> I think it's good if you go into a study like "I am creating data to be shared," that may change your approach you may make sure that you're like being extremely methodological that there's nothing that you do that isn't like written down and well documented so that later there is this very well documented approach. I don't know if you've ever gone the National Center for Health Statistics website all of those surveys are intended to be shared right it's public use data when you look through their documentation, I mean it is just incredibly detailed… (P3)

Researchers planned from early stages in the project to deposit data, revealing that fitting data to the repository standards was becoming an increasingly taken-for-granted part of their research workflows. When faculty anticipated that they would need to eventually deposit data, they tended to create systems from the start of the project for data management. They were more cognizant that the data would need to be described, their provenance chain documented, and the scripts well-commented for future interpretation. This anticipatory attitude, and the consequent systems reflecting the orientation to the data's eventual reuse. For example, P4 (Study 1) reported that researchers are anticipating what repositories require from them early in the research cycle. In response to the question: **"**when do you know something is ready for submission to the data repository?" a genomics faculty member in a public R1 institution replied:

> *I wouldn't submit until it's what they want, usually what they want…they would like to see the negative data too. And so that didn't necessarily go in. They want to see what you've screened and what was negative.* (P4)

In genomics, depositing data was standard practice, well-established as part of research workflows, given funding pressures and publisher agreements that require data sharing. A theme that emerged in the comparison of the two was the perceived legitimacy of qualitative (versus quantitative) data. In the social sciences, there was often a need to justify the qualitative methods they used in their research. However, genomics researchers did not bring up the issue of the legitimacy of methods or data. Granted, participants expressed concerns about the "curse of high-dimensionality," that is, the high probability of spurious correlations in big data – e.g., P1, a tenured faculty in reproductive evolution, was skeptical of researchers who entered their data analysis without a clear set of research questions, accusing that researcher of "fishing" for

results. Nonetheless, there are well-documented institutional measures in place to account for such concerns, e.g., registering hypotheses prior to analysis (Mellor, 2017; Nosek et al., 2018).

In the social sciences, however, the legitimacy of qualitative data and methods were questioned, both from within and outside the participants' disciplines and sub-disciplines. For example, qualitative data analysis (e.g., of interviews) was seen as not "rigorous" enough to merit data validity. P4 (Study 3), a tenured faculty in public health, used ICSPR to legitimize her qualitative data. She saw the ICPSR as not only a way to share her data for potential reuse but also to enhance the trustworthiness of the dataset, saying:

> *Well, how many people can say their data set is with ICPSR? When you think about University of Michigan, they have a lot of really good data programs and certificates… I think this is a game changer for those of us who do qualitative work because, I mean quantitative researchers do it.* (P4)

For P4, ICSPR served an important legitimizing function for her qualitative data and methods. The repository was perceived by P4 as lending the institutional authority of the University of Michigan to her qualitative approach, not least because of the "really good data programs and certificates." (P4, above). P4 and other qualitative researchers in the study described how the data they deposited gained credibility from the reputability of the repository, not least because of its accompanying certification. P4 and other researchers in Study 3 found themselves justifying their work to others to legitimize their methods. While the questions of legitimacy of quantitative and qualitative paradigms likely will not be resolved, participants across the three studies demonstrated the role of institutions such as data repositories for establishing trust and legitimacy of data. The disciplinary differences, then, also come down to more rather distinctions within sub-disciplines of research fields according to ongoing debates about qualitative vs quantitative methods, and data analysis approaches.

7.1.4 Why is it hard to institutionalize RDM?

In the interviews, participants described the advantages and disadvantages to institutionalizing data management. First, they reported it is hard to institutionalize an ever-changing tool or process. With institutionalization comes routines that are well-established, taken-for-granted, and which eliminate uncertainty (DiMaggio & Powell, 2000). It takes time for the tools and practices of RDM to become well-established and normative. So, when the tools and processes for RDM often change, it is difficulty to concretize them enough such that they are taken-for-granted. For example, P2 (Study 3) is a tenured professor in epidemiology and aging studies who explain that having the same data management approaches across project was challenging "because universities are also changing their systems all the time" (P2). Similarly, P5 (Study 3), a pre-tenure research faculty in criminology described how the systems would change, or there might be delays in subject recruitment, data collection, or another aspect of the process, complicating the research teams' ability to institute processes consistently:

> *You say like somebody said that you know we're going to get the data from this hospital system and then, when we get it it's like so messy that it takes us two months to get into any kind of shape that we could then do analysis on it So it really just varies by project, but I would say there's always something like I can't tell you here's what it always is it's just there's always some kind of delay, for whatever reason. And it's usually around like receiving data or data collecting data or it's like "Oh well, we had this plan in place to talk to young people in the sex trade and it's just going slower than we thought recruitment is going slower than we thought* (P5)

The unexpected aspects of the work P5 reported make it hard to institutionalize (e.g., there are unexpected delays in receiving the data or collecting the data). One of these unexpected aspects is AI-Human collaboration for anonymization. Specifically, it is hard to institutionalize a managed process[13] for RDM when the process involves an unresolved technical problem. One

[13] "Managed process" is language drawn from the capability maturity model developed in the software development community to refer to process optimization (it is also what Crowston & Qin's refer to in their capability maturity model (CMM) for RDM (2011).

site where this came up frequently was the anonymization and de-identification of qualitative data such as interview transcripts. This is a technical problem; it cannot be fully automated, even though participants being part of initiatives to develop semi-automated solutions to reducing the work of de-identifying transcripts in qualitative research. Specifically, several social science participants (Study 3) reported being part of the Washington University in Saint Louis's (WUSTL) initiative to test a natural language de-identification software product (Gupta et al., 2021; Mozersky et al., 2021). However, the software's performance was not 100% accurate at de-identifying transcripts, a rate of false positives and negatives which has a large social cost. In the case of false negatives, the transcript is de-identified in a place it should not be, sacrificing the details of the data by unnecessarily removing contextual information. On the other hand, false negatives are a worrisome case that would breach confidentiality were they not caught by the researchers to identify where machine-learning anonymization algorithms went wrong.

In this instance, creating structures and routines to institutionalize the automated system would be difficult, because working with the anonymization software remains a highly local, context-dependent procedure. For example, P1, a bioethicist described how working with the anonymization AI (WUSTL pilot project) to prepare her data for deposit required deep knowledge of the relevant field. In her case, she had to correct for the algorithm's oversight – it was not even a 'mistake,' or 'misclassification,' per se, because the work of anonymization is anticipatory – where the software did not anonymize for gender because social workers are 85% female. If she did not anonymize for gender, it would be relatively easy to triangulate location information and gender to identify a study participant.

> *Gender ended up being a thing because I interviewed social workers. Most social workers are 80% female in the United States. But there are male social workers. I actually anonymized gender…because it's so rare that there's actually a male social worker. That [would] be identifiable. I think you'd have to work really hard, but it's not like the software will anonymize gender…It's not going to change "he" or "she," or put*

> *brackets around it, or just say "social worker" instead. I had to do all of that myself.* (P1)

As with the case of P1, it would not be clear what institutionalizing the process would look like because of the contingencies of transcript de-identification. What would that look like? It might be a budget for a research data manager. The data manager would have to be embedded with the team but require a great deal of training and even a dual role as the data collector, researcher, and the person who deposits. As P1 elaborated: "having a proxy to help would not really work." In CMM for RDM terms, the "ability to perform" requires resources like personnel to help with deposit, e.g., such as a "proxy," is not clear, here. Other participants reported its important to anonymize other discipline-specific items (e.g., P6/P7 incarceration status).

In addition, there are advantages of not having protocols and procedures, as well. Bespoke projects require somewhat bespoke solutions. The findings from across the three studies suggested that part of this is the sensemaking aspect of data management is a big part of why not employing protocols and procedures were advantageous. The "messiness" of the project lent itself well to sensemaking by researchers, as (Sawyer et al., 2015), as well as allowing for coordination to occur among group members. As P7, a tenured criminologist, emphasized, it can be useful to keep the process unstructured:

> *By virtue of this project and not falling into a project world of protocols and procedures were set but having to confront them and create solutions. And you know, we had no money to send him down to look at blaze as a software package, but we sent him in June down to do their five-day training and so he we looked at salesforce as well. Which is perfect for track and follow ups and the like, but it just didn't fit the timing for our project.* (P7)

What P7 refers to is the benefits, and necessities, of flexibility. Once they had selected a secure software system to manage their data, P7's project team started to develop procedural guidelines. However, before that, they were resistance to prescribing a single way to do RDM. Across the three studies in this dissertation, participants also reported resistance to concretely defining procedures where there is lack of agreement about a 'best way" or even "best practices"

e.g., for ethical treatment of subjects, consent, and in genomics case, how to help students learn RDM, and using digital lab handbooks.

7.1.5 Defining data quality: more than accurate data

An emergent theme in the interviews with participants – in genomics and the social sciences – was data quality. When discussing data management activities across the lifecycle, including collecting data, documenting data, and preparing data for deposit, participants defined data quality beyond mere accuracy. For genomics researchers and social scientists, it was important to make sure the instrument used to collect data was reliable. The results, if not, were pervasive if not caught early in the data management pipeline. For example, as P11 (Study 1), a genomics research faculty studying psychiatric disease described:

> *We analyzed the data, we see a very strikingly, huge difference, which surprised us completely. We never had seen this kind of striking difference before treatment and after treatment. Then, after analyzing that data, we learn it's just technical artifacts. The commercial chip has a quantification bias. So that's a technically confounding, actually screw up your whole data. I just have to redesign the whole experiment. It's a problem nobody ever talked about. So far, people used that chip for even four or five years now. Millions of people never realized. Then the result analysis will be screwed by those kinds of artifacts.* (P11)

The issue of the artifact in the data were overlooked for years, and led to wasted resources (e.g., the time of the researchers in writing up the spurious results, the expenditures for computing time). If using a core facility to produce or process data, the checking for quality was not as relevant. For example, a genomics faculty (P12, Study 2) trusted the university's sequencing core to produce the data accurately. A social epidemiologist (P14, Study 3) described how she defined the parameters and quality check for the core facility – specifically, the "recharge" research center – to ensure data quality at a point further upstream than its production. In the words of P14, the recharge center processes the data in a trustworthy manner. Yet P14 still performs data quality checks to ensure her instructions were followed:

> *I create specifications, like, I'll write a paragraph or a list. Then they take that structure, and they highlight the specific files that the information is going to come from and the specific variable names that are needed*

> *and say exactly what to do… [The research computing center] has almost like a manual. There'd be SOPs, the standard practices where they I'm sure that they have standards that they follow, including file naming conventions, storage. Some of this is dictated by the data use agreement that you have to have in place with the entity that's letting you use the data. They have – especially CMS [Medicare] data – they have requirements that you* must *meet and they're super specific about the encryption on the files and the security of the server…* (P14)

P14 was assured of the quality of the data in part because she relies on the trustworthiness of the research computing center having standard operating procedures (SOP) and professional best practices for dealing with Medicare data. However, assuring data quality often required faculty to intervene and set up checkpoints in the data management and analysis process to ensure their students were properly collecting, analyzing, and managing research data. Collaborators got involved in data quality, in terms of accuracy, means more than just the grammar and spelling, but also if the questions were answered and asked correctly. For example, do the students know pop culture references and French-Canadian accents (P2, Study 3)? Were follow-up probes questions asked (P1, Study 3)? P2 highlighted how they had to guide some students in the data collection to ensure no questions were skipped in the protocol:

> *In 2019 there were a few students who probably skipped questions about like so one of the questions was like do you think sexual and one of the questions was like is there a couple questions I think probably uncomfortable more than like students are already uncomfortable interviewing somebody, and so I do feel like a couple of students like just those and so that's not good for analysis right?* (P2)

As well, the representativeness of the data of a target population was a key issue for P13, a faculty in behavioral health, who saw data quality in terms of whether it was inclusive: "Previously, it was predominantly white respondents, and so it was problematic… That's where a lot of the data quality matters" (P13). In addition to data collection and representation, faculty emphasized the stage of data analysis as a point for qualitative data quality. P9, a tenured faculty in health services and substance abuse, described how comparative coding was a technique vulnerable to data quality concerns. In P9's words:

> *The point is, if you have extremely nuanced codes, to the point that you only have two excerpts in a code well when you export the excerpts and then try to summarize them and look for consistencies and inconsistency…. That's what I mean by quality in terms of like consensus coding. Now on the content analysis side of policies, there should be a generally pretty clear consensus of what the law is saying. With that I tend to use again an inter-rater reliability test. With student I say let's practice. That means that the very beginning, we will do some consensus coding just to make sure we're all on the same page about what the codes mean, we have clearly defined code clear definition. But then eventually we switched to an inter-rater reliability test. I want to see usually at least like a point eight Kappa.* (P9)

P9 had a systematic approach for judging the data quality, as well as quantitatively measuring the inter-rater reliability when conducting content analysis of the data. Other connotations of data quality the participants highlighted as important included ethical issues in research data management and data deposit (e.g., P6, Study 1).

7.2 Contributions and Implications of the Dissertation

This dissertation contributes to the field of information science and technology. The research and practical outcomes contribute to the specific area of data management (RDM). The contributions include advancing methods, theory, and practice, addressed in the next sections.

7.2.1 Methodological Contributions

The methodological contribution of the study is threefold: qualitative approach to the institutionalization of RDM, adding to virtual methods in RDM (i.e., using a critical incident technique via Zoom), and developing an analytic tool to measure institutionalization of RDM and articulation.

First, this study used qualitative research methods (e.g., interviews and a grounded theory-inspired approach) to address questions of data management and deposit practices. Prior studies of data management and data deposit in information science and library and information science (LIS) have largely employed quantitative approaches such as surveys (e.g., of faculty or librarians) or science of science approaches (e.g., bibliometric, scientometric) methods to investigate data practices and attitudes. These are valuable approaches that have led to revelations about the factors influencing researchers' attitudes about RDM. Yet, these studies do

not focus on the relationship of attitudes and practices leading to data deposit. Granted, in science studies ample prior research has used the ethnographic and hermeneutic approaches often associated with the study of scientific practices (e.g., laboratory studies such as those of Collins et al., 2003; Latour & Woolgar, 1986; Neang et al., 2020; Pinel et al., 2020). In fact, many argue laboratory studies employing ethnographic approaches are more appropriate fitting for investigating research practice, routines, and institutional logics. Yet, these studies have not traced the practices of data deposit to open research repositories. Organizations will increasingly develop guidelines for RDM; studying this institutionalization of RDM requires methods appropriate to the study of institutions, which are served well by qualitative methods, useful for uncovering human perceptions, beliefs, and attitudes (Creswell & Poth, 2016).

The field of information science and technology studies of RDM can benefit from employing qualitative approaches to examining institutional aspects of RDM. For example, researchers who aim to promote long-term data sustainability need to focus on advancing our understanding of the institutions—policies, practices, norms, and beliefs — that lead to effective data curation and stewardship. To advance this goal, approaches like ethnography can assist in articulating the institutional logics that are associated with higher data deposit. Qualitative approaches can be of help, as well, to uncover institutional norms and behaviors of organizations that steward data (e.g., GenBank) and individuals (e.g., researchers) who instantiating RDM policies (e.g., NSF data management plan mandates).

While these studies do not unilaterally need to use qualitative approaches, they stand to benefit them because of their ability to surface the social phenomenon which are not possible to observe directly, such as the cultural-cognitive and normative pillars of institutions. For example, indicators of the cultural cognitive pillar of institutions include “common beliefs” and “shared

logics of action" (W. Scott, 2008). This dissertation adds to the growing of work on the institutionalization of RDM, adding the innovative approach of qualitative inquiry to scientometric and quantitative studies.

Second, the Study 2 and Study 3 of this dissertation used virtual synchronous interviews (Hine, 2005) and a cultural probe (i.e., dataset metadata records shown to the participant). Increasingly, research has employed virtual methods (e.g., data collection using zoom, Skype, Google Jam boards). Using virtual methods is an approach that will continue to grow, given the utility of the features for qualitative research analysis (e.g., audio recording, video recording, whiteboard features, transcription features). In Study 2, the video interviews assisted with the collection of data during the COVID-19 pandemic and can be a model to be employed by other researchers in the study of data deposit by researchers to online repositories.

Third and finally, the methods to address questions of Study 3 led to the development of an analytic tool for measuring the institutionalization of RDM. The core measures came from an approach to measuring institutionalization of data management developed by Crowston & Qin (2011): a capability maturity model for RDM (CMM for RDM). However, the CMM for RDM is a long rubric with many indicators of various levels of maturity. Specifically, the assessment rubric has approximately 400 measures (4 areas of maturity x approximately 4 items each x 5 levels of maturity x 5 process areas). The total comes to 378 items, which makes the CMM for RDM a cumbersome for scientists to use as an assessment tool in their own labs or researchers studying the maturity of RDM to employ as an operational measure of RDM institutionalization.

In addition, measures for "articulation" in response to the institutional were defined. In Study 3, I developed a preliminary deductive framework that brings together Fujimura's levels in the articulation with measures of institutionalization. Core measures of the articulation work

done to manage RDM institutionalization are defined (shown in **Table 16,** Ch. 6). The analytic measure of articulation contributes to what has been a theoretical description of articulation activities (e.g., "alignment" "work that makes work *work*"). The vagueness of some of these descriptions can make it challenging to measure. By adapting Fujimura's articulation framework, this study contributes a deductive model of articulation work which can be used in future research to enable comparisons across studies of articulation in RDM.

7.2.2 Theoretical Contributions

This dissertation contributes to the existing foundation of theoretical knowledge through its conceptualization of research data management as "articulation" and by bridging the gap between the empirical and theoretical research in this increasingly impactful area. The study also intersects neo-institutional theory and capability maturity models for research data management (CMM for RDM) together to cross-pollinate concepts and techniques between two fields.

Study 2 developed the concept of "data articulations" to draw attention to the practical work faculty do to prepare data for deposit. This concept, coined as *datarticulations*, describes the iterative activities that researchers perform to reconstitute the data deposit workflows. Study 3 elaborated the datarticulations framework and developed into a conceptual analytic that assists with understanding the labor supporting CI-enabled science by attending to the 'various and variably configured conditions of alignment of the many levels of work organization' (Fujimura, 1987: p. 283). RDM scholars can use the concept of *datarticulations* to understand how researchers produce and share research data. Science policymakers to design interventions for supporting faculty development in data management and genomics data workflows.

Further, the CMM for RDM developed by Crowston & Qin, and other applications of the model or derivates of their assessment tool, do not link the maturity of a managed process to the

institutional literature. In this study, there was a formal link made between (neo)institutional theory and the capability-maturity model for RDM. What this does is develop a bridge between the ability to measure RDM maturity and the literature on understanding how the processes of increasing maturity become take-for-granted, that is, become *institutionalized.* Not only can we measure to what extent data management is mature, but we can also say why and how it came to be, enabling recommendations to researchers and policymakers who want to increase the maturity of data stewardship and long-term research data sustainability in academic research labs. Future work can build on this analytic tool and core measures of RDM institutionalization.

7.2.3 Practical Contributions

The study findings have some practical implications of interest to decision makers (e.g., science policymakers, data repository administrators), and funding agencies to promote the long-term sustainability of research data and provide data services that fit the needs of researchers in genomics and the social sciences.

First, bridging the CMM for RDM with institutional literature makes an important practical contribution because it assists in our ability to assess what fields need greater maturity, but also can inform recommendations for improving RDM at multiple levels. For example, scientists and their lab groups can use the tool as evidence of RDM maturity to funders. The momentum for promoting long-term data sustainability for organizations as diverse as academic research institutes in biology and genomics (e.g., NCBI institutes, Bio5 and Biosphere2 affiliated with the University of Arizona) and in the social sciences (e.g., Dataverse at Harvard University and the Interuniversity Consortium for Political and Social Science Research (ICPSR).
Second, the artifacts gathered are useful for researchers who seek to codify best practices for RDM in their labs. In study 3, the social science participants reported their desire for templates

or examples for managing their lab research data. For example, they lacked written guidance for their students to analyze data. Instead, participants would verbally instruct students in how to analyze the data. Participants complained that orienting new students verbally was redundant; every year, the faculty had to explain the analysis process again. To address this redundancy, some participants developed to fill the gap in the lack of documentation to formalize and codify procedures to guide students in data analysis. Study 3 showed that there is value in sharing data and project management documents among researchers, and that some researchers already informally circulate such documents among colleagues. Therefore, to mitigate redundancy issues, a repository or collective database for such documents and a community of practice around them, would help researchers to manage their data by drawing from the solutions developed by others in their fields.

Third, the studies identified areas of policy where researchers struggled most. That is, the studies illuminated possibly locations for interventions by policymakers or institutional leaders to support scientific data management. This dissertation adopted the lenses of articulation in relation to institution; As such, the studies identified key consequences of institutionalization of data work, and can inform recommendations for science policymakers, scientists, and information science professionals who support scientists (e.g., metadata librarians). As a result, the study findings address what types of data and data management practices should be more institutionalized and whether we should institutionalize the articulation work. The study findings suggest ways in which genomic and social science data management institutionalization may be generalized to other fields. For example, the studies identified the aspects of data management that make it hard to institutionalize them as “managed processes.” These include unresolved technical problem, unexpected issues that require flexibility, or fast-moving organizational

change (see section "Why is it hard to institutionalize RDM?" above). These findings have implications for the use of artificial intelligence (AI) in research data management workflows. For example, cleaning and deidentification of data cannot be fully automated yet because of the high error rate relative to the risk of breaching confidentiality.

Researchers resisted the data deposit mandates because they posed a challenge to researchers who did not have the time, human resources, or budget to manage their data. The comparison between genomics and social science researchers also illuminated an important difference: the ease of and resource allocation/support for data deposit. The challenge for social science data was often deidentification and the lack of budget and personnel for RDM. A notable exception was the social science researchers who participated in the Washington University at St. Louis (WUSTL) qualitative data and machine learning software project. The insights into how policy and practice intersect are important for policy assessment. They can inform evidence-based data policy such as those of funding agencies and repositories to create mechanisms for supporting data curation. AI in data curation is a field that will grow, and the development of automatic tools for deriving metadata, documenting data, and verifying the integrity of data can assist with advancing the field by reducing workload of scientists in managing and depositing datasets.

7.3 Limitations

As with all research, there are limitations to the study. The limitations are primarily related to the selected populations, method, and the COVID-19 pandemic impacts on data collection for Study 3. The limitations have implications for generalizability, but also point to opportunities for future research.

First, a limitation of the study was the population sample and its generalizability. The study included genomics researchers and social scientists from various sub-disciplines, which is

not as focused a sample to inform the codes. However, there is a chance that the codes that came up in the interviews did not capture other factors due the small number of subjects and the subset of sub-disciplines within genomics and the social sciences. Further, the studies did not use stratified sampling. As a result, the findings of the interviews treat genomics and the social sciences as broad fields. Hence, the study is limited in its capacity to draw conclusions about the sub-disciplines because comparisons between these specializations would require a) a deeper examination of the norms, traditions, and history of the fields involved, b) a revised methodological orientation, and c) a reconsideration of the interdisciplinarity of the fields included. Notwithstanding these limitations, this research proceeded on the premise that it is critical at this stage in the research on RDM to explore the disciplinary differences and similarities to improve understanding of RDM, especially given that the findings suggest differences in sub-disciplines.

Second, the deductive model derived from the CMM for RDM was only a subset of indicators of maturity of RDM. A subset of core measures was selected as a preliminary method for measuring the institutionalization of data management and deposit. These core measures included a) the presence or absence of data documentation b) the presence or absence of structured workflows c) the extent of regulative or governance policies for RDM d) the presence or absence of file naming conventions. Although deriving this subset of core measures was useful because it reduced the CMM for RDM from approximately 400 items to 9 core measures, it was limited in its scope and the granularity of the items. As a result, it was difficult to show associations between the 9 measures and the resultant analysis section of Study 3 was more interpretive. Work remains to be done to test the efficacy of the core measures selected to represent and operationalize the institutionalization of RDM.

Third, although multiple rounds were taken in the coding process, it was conducted solely by the author. Although single-annotator data analysis is a valid method (McDonald et al., 2019), it would be useful to triangulate the analysis with other coders. Using multiple coders would have enabled the study to perform inter-coder reliability demonstrate the strength of the factors surfaced in the interviews. To mitigate for this in part, Study 2 was conducted with researchers and the document was in consultation with the dissertation committee, especially the advisor, over the course of its development.

Fourth, the focus of the study was on successful data deposit. This focus was motivated by the research questions that centered on understanding how faculty organize their labs to enable data management and deposit. Although the interviews did capture some of the instances of failed data deposit through the faculty narratives, the scope of the research was defined to primarily delve into cases where deposit to a research repository was achieved. The limitation of this focus is that it precluded an assurance of the validity of the factors identified as associated with mature RDM and data deposit. Further, by focus on successful deposit, the research was not able to gain a deeper understanding of the barriers to deposit, which could have valuable practical implication for helping scientists to share their data in research repositories.

Finally, the data collection with social scientist was initially designed as in-person collection. The COVID-19 pandemic led to video conferencing for data collection instead. The limitation of this was threefold: first, it was difficulty to glean a comparable level of detail about the environment which the faculty was embedded in. In the Study 1 and Study 2, I visited the labs, buildings, offices, and campus of the researchers and took pictures. The faculty also were able to refer to items onscreen (e.g., we looked up data deposit to Mouse Genome Informatics (MGI) database), and through these interactions, snowball sampling was facilitated because the

faculty member and I would run into another person in a laboratory down the hall. The pandemic also limited the participation of researchers who are in marginalized groups or with families who declined to participate (that is, they directly declined citing this as the reason, or they potentially declined). Finally, the participants were dealing with stress and different research pressures and research environments than pre-pandemic.

Although efforts were taken to mitigate the limitations of these methods, I recognize the constraints which should considered for evaluating of the study results. Notwithstanding these limitations, the study has applicability outside of the initial population and can extend to biosciences faculty in R1 academic institutions who deposit data and social scientists in data-intensive disciplines. Because the purpose of the study was to explore the ways that researchers organize their work such that they enable data deposit, these limitations posed some constraints but also pointed to promising directions for future research. Some of these opportunities for further study are outlined in next section.

7.4 Opportunities for Future Research

Given that studies concerning how and why researchers manage and deposit their research data in an increasingly institutionalized research environments for RDM are relatively new, several areas for examining this phenomenon remain untapped. Based on the findings and limitations of this dissertation research, I suggest four directions for future research in what follows: The future research opportunities are directions that can make important methodological, theoretical, and practical contributions by extending on building on this study.

For one, this study encourages future research on RDM to consider applying and extending the conceptual framework developed in Study 2 and the operational measures for institutionalizing RDM in Study 3. The “articulation” framework can be used by RDM workflow

design researchers to identify the existing technical challenges researchers have when submitting data to a research repository. More broadly, looking at articulations can also help policymakers to develop researcher-responsive incentives and evidence-based policy (e.g., data mandates) that support the researchers. Importantly, these concepts can assist researchers to see how researchers are filling the gap in system design and science policy to make data deposit more efficient and effective. Articulation work occurs where there is an institutional vacuum, that is, a lack of capital to support RDM – whether human capital, financial, or informational – or other resources. As well, research building on this can add factors and facets which may have been overlooked in the qualitative approach to surfacing factors. In particular, the use of the *datarticulations* framework in a sub-discipline of a scientific field would be especially fruitful to elaborate the theory.

Based on this finding, a second suggestion for future research is to apply the conceptual framework to sub-disciplines within the social sciences. To address the limitation of this dissertation that did not use a stratified sampling technique, future work can use the approach to ensure the significant minimum participants count for each (sub)discipline. As part of this, the results would compare the respective subdisciplines in terms of the idiosyncrasies of their data management and data deposit practices. The value of this contextualization is to identify, and potentially control for, the crucial points for scientists' data management practices. and as a result, develop tailored strategies for each subdiscipline. As well, complementary methods for comparing groups could leverage the qualitative methods such as ethnography, "making tea" (schraefel & Dix, 2009), focus groups, and/or design probes to assess the full range of practices and general norms scientists follow for managing and depositing data to repositories. Similar approaches to within and between group comparisons could also be adopted to differentiate the

data practices by using other methods, e.g., to explore the extent to which policy factors associated with institutionalization of RDM impacts individual data practices, if there is a generational divide in use of repositories, or if RDM training (e.g., graduate curricula) impacts the maturity of data management in a researcher's lab.

Third, the study findings suggested a distributed workforce in areas of higher institutionalization of RDM. For example, genomics researchers' workflows involved a distribution of the workflow among a variety of actors with specialized knowledge. Future work can look at coordination work given the need for coordination work in a distributed workforce and workflow. This extends into peripheral areas to information science and technologies and LIS studies, such as that of the Future of Work. The distribution of labor across core facilities (e.g., Molecular core), material suppliers (e.g., for reagents), sequencing companies (e.g., genewiz), software developers (e.g., BaseSpace), data analysis contractors, and the academic research lab has implications for questions of how formalizing practices, i.e., institutionalization, has implications for creative, flexible workflows. The results of this study showed that researchers do employ creative methods to account for tightened budgets or to manage unexpected circumstances such as using dental record-keeping software as a non-cloud based RDM solution (e.g., a criminologist, Study 3) and glue guns to set up an otherwise expensive experiment (e.g., genomics, Study 1).

Researchers in areas outside of information science can bring theories to these finding to extent them such as the theories of anticipation work (e.g., science and technology studies (STS)), and tool appropriation (e.g., Computer-Supported Cooperative Work (CSCW)). Researchers in neo-institutional theory can apply the theory to questions of a distributed workforce to further develop the initial link made in this dissertation between the CMM for

RDM and institutional logics, e.g., to develop a deeper understanding of how the institutional logics constituting RDM practices are formed and how they diffuse along the research process, especially given the distributed nature of the research processes. Future work can explore this rhizomatic (Deleuze & Guattari, 1988) relationship between research design, RDM, and long-term research data sustainability.

Fourth, this dissertation takes a closer look at what is the composition of "top-down and bottom-up" data management "maturity" (using the CMM for SDM and other models). Top-down indicates a widespread, institutional norm. Bottom-up insists on the "home grown, "fill in the blanks", and faculty or student driven instantiation of data management systems. Articulation work foregrounds tradeoffs made between the global and the local. Future work can examine questions of how genomics scientists negotiate to "make the global local," setting up their labs to meet the standardized and often institutionalized demands of data repositories. As such, future research questions in this area can also examine RDM guidelines as an *adoption of innovation* story. Such questions can ask: How do faculty *inhabit* and *domesticate* RDM institutionalization? What infrastructures do they draw from and what social or organizational mechanisms do develop or use to tailor the mandates and policy to their actual work practices? Relatedly, in comparing various disciplines with respect to institutionalization and articulation, future studies can move beyond the theory adopted by this study, which was an institutional perspective. Future work can also conceptualize the study samples in terms of "high-paradigm, low-paradigm" fields to analyze how the maturity of a discipline (in a Kuhnian-paradigm sense) is related to its maturity of data management. For example, future research could apply the "scientific paradigm" lens to test a hypothesis that the maturity of a discipline or fields' methods

and convergence on basic axioms are positively related to the maturity of Its data management (where the *maturity of RDM* is measured by the CMM for RDM model (Crowston & Qin, 2011).

Finally, future research can take design approaches as well, such as user experience design. The research findings surfaced multiple instances where faculty were incorporating artificial intelligence (AI) into their workflows for data management. This development leads to opportunities for addressing the question of RDM and data sustainability in the context of AI-enabled data curation and management; the practical application of AI for extracting *paradata.* Outcomes can include prototypes designed to help genomics researchers and biologists to described and share their data in a FAIR way. These prototypes enable their data and associated publications to be machine-readable, and thus "readily harvested for large scale cross group analyses" (Cui et al., 2018). Such prototypes offer a site for collaboration amongst researchers.

7.5 Conclusions

The overarching question of the study was how does institutionalization of data management and deposit impacts long-term research data sustainability? In specific, four research questions (RQs) guided the study: 1) What are the experiences of data deposit for genomics faculty? 2) What faculty data practices make data deposit 'do-able'? 3) What institutional factors are associated with "articulation" of data management and deposit? 4) What are impacts of "articulation" on long-term research data sustainability?

Following a sequential qualitative approach, this dissertation explored the institutionalization of faculty research data practices, attitudes, and perceptions in the context of managing and depositing research data to an open research data repository. In three consecutive studies, the dissertation addressed a gap in the conceptual and empirical understanding of the impacts of institutionalization on long-term data sustainability. Study 1 explored the data

practices of research-active faculty in genomics to surface the practices involved in deposit research data. Study 2 theorized data deposit as “articulation work,” and adapted Joan Fujimura’s theory of articulating alignment in data management and deposit. Study 3 applied the conceptual framework to determine its fit in a broader population of social science faculty, and to surface the factors associated with “articulation work” in data management and deposit.

The studies contribute to the dissertation’s goal to a) identify the factors associated with ‘articulating data institutionalization’ in big science and little science fields; and b) identify the impacts of articulation on long-term research data sustainability. Genomics represents a ‘big science’ field with mature data institutionalization (e.g., data deposit to GenBank). Sociology and political sciences represent a ‘little science’ field with less institutionalized data deposit infrastructure (data deposit, e.g., to ICPSR). The study design focuses on data deposit to data repositories because they are a signpost of data institutionalization and rich site to study how faculty adapt data policy to local circumstances through articulation activities.

Study 1 found RDM has become more institutionalized, and that faculty often must reorganize their workflows to accommodate directives, mandates, and cultural pressures to deposit data. Study 2 built on these findings, developing a model to explain the process of data deposit. This study found faculty engage in articulation work in response to institutional pressures to deposit data. The processes include setting checkpoints or ‘thresholds’ for data deposit to ensure data is suitable and contingencies met. We also found some outcomes of articulation are aligned with the goals of long-term research data sustainability including ensuring data quality, integrity, and completeness. Study 3 found that articulation can create opportunities for flexibility but can undermine sustainability. Templates for RDM were created in the vacuum left by lack of institutional support.

Contrasting high- and low-institutionalization research data management contexts to identify factors that promote or inhibit long-term research data sustainability. Using mixed methods including semi-structured interviews and document analysis, I argue data management work is not extraneous but central to requirements assessments for process improvement and evidence-based data policy. The approach of data "articulation" challenges the long-standing paradigm which considers later stages of the data cycle as key sites for engaging the sustainability dilemma. This approach brings institutions and practice back into the analysis, opening new directions for empirical studies of the work behind data curation.

Broader implications of the findings are that when we require scientists to perform data management, e.g., describing data, and we have institutionalized those data processes, and then researchers anticipate and shape how they design their research going forward. In other words, the institutionalization of data description, organization, and storage then begin to influence how we design our research. The expectation to deposit influences how they design research. We say this in social science researchers, in that they anticipated requirements and institutions get embedded in the workflow. Future work can build on and extend these findings and the conceptual framework developed to explain the data deposit practices, and the institutional supports to promote long-term research data sustainability.

REFERENCES

Abrams, S., Cruse, P., & Kunze, J. (2009). *Permanent Objects, Disposable Systems*.

Ahlin, E. M. (2020). Forced Sexual Victimization Among Youth in Custody: Do Risk Factors Vary by Gender and Perpetrator? *The Prison Journal*, *100*(2), 151–172.

Akers, K. G., & Doty, J. (2013). *Disciplinary differences in faculty research data management practices and perspectives*.

Alexander, S., & Gray, J. (2006). 2020 Computing: Science in an exponential world. *Nature*, *440*, 413–414.

Alidina, H. M., Fisher, D., Stienback, C., Ferdana, Z., Lombana, A., & Huettmann, F. (2008). Assessing and managing data. *Marxan Good Practices Handbook*, 14–20.

Allen, L., O'Connell, A., & Kiermer, V. (2019). How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship. *Learned Publishing*, *32*(1), 71–74.

Allen, L., Scott, J., Brand, A., Hlava, M., & Altman, M. (2014). Publishing: Credit where credit is due. *Nature News*, *508*(7496), 312. https://doi.org/10.1038/508312a

Alperin, J. P., Schimanski, L. A., La, M., Niles, M. T., & McKiernan, E. C. (2020). The value of data and other non-traditional scholarly outputs in academic review, promotion, and tenure in Canada and the United States. *Open Handbook of Linguistic Data Management*.

Álvarez-Machancoses, Ó., DeAndrés Galiana, E. J., Cernea, A., Fernández de la Viña, J., & Fernández-Martínez, J. L. (2020). On the Role of Artificial Intelligence in Genomics to Enhance Precision Medicine. *Pharmacogenomics and Personalized Medicine*, *13*, 105–119. https://doi.org/10.2147/PGPM.S205082

Ankeny, R. A., & Leonelli, S. (2015). Valuing data in postgenomic biology: How data donation and curation practices challenge the scientific publication system. *Postgenomics: Perspectives on Biology after the Genome*, *126*.

Antes, A. L., Walsh, H. A., Strait, M., Hudson-Vitale, C. R., & DuBois, J. M. (2018). Examining data repository guidelines for qualitative data sharing. *Journal of Empirical Research on Human Research Ethics*, *13*(1), 61–73.

Arias, J. J., Pham-Kanter, G., & Campbell, E. G. (2015). The growth and gaps of genetic data sharing policies in the United States. *Journal of Law and the Biosciences*, *2*(1), 56–68.

Armstrong, S. J., Allinson, C. W., & Hayes, J. (2002). Formal Mentoring Systems: An Examination of the Effects of Mentor/Protégé Cognitive Styles on the Mentoring Process. *Journal of Management Studies*, *39*(8), 1111–1137. https://doi.org/10.1111/1467-6486.00326

Atkins, D. (2003). *Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure*.

Atkinson, R., & Flint, J. (2001). Accessing hidden and hard-to-reach populations: Snowball research strategies. *Social Research Update*, *33*(1), 1–4.

Austin, C. C., Brown, S., Humphrey, C., Leahey, A., Webster, P., & Fong, N. (2015). Research Data Repositories: Review of Current Features, Gap Analysis, and Recommendations for Minimum Requirements. *IASSIST Quarterly*, *39*(4).

Ayers, E. L. (2004). Doing scholarship on the web: 10 years of triumphs and a disappointment. *The Chronicle of Higher Education*, *50*(21), B24.

Baker, K. S., & Bowker, G. C. (2007). Information ecology: Open system environment for data, memories, and knowing. *Journal of Intelligent Information Systems*, *29*(1), 127–144. https://doi.org/10.1007/s10844-006-0035-7

Baker, M. (2016). Reproducibility crisis. *Nature*, *533*(26), 353–366.

Bala, A., & Gupta, B. M. (2010). Research Activities in Biochemistry, Genetics and Molecular Biology during 1998-2007 in India: A Scientometric Analysis. *DESIDOC Journal of Library & Information Technology*, *30*(1), 3–14.

Barber, B. (1952). *Science and the Social Order, with a forward by Robert K*. Merton. Glencoe, IL.: Free Press.

Bardach, E., & Patashnik, E. M. (2019). *A practical guide for policy analysis: The eightfold path to more effective problem solving*. CQ press.

Barley, S. R. (1986). Technology as an occasion for structuring: Evidence from observations of CT scanners and the social order of radiology departments. *Administrative Science Quarterly*, 78–108.

Barley, S. R., & Tolbert, P. S. (1997). *Institutionalization and Structuration: Studying the Links between Action and Institution*. https://doi.org/10.1177/017084069701800106

Bärmark, J., & Wallén, G. (1980). The development of an interdisciplinary project. In *The social process of scientific investigation* (pp. 221–235). Springer.

Bates, J., Lin, Y.-W., & Goodale, P. (2016). Data journeys: Capturing the socio-material constitution of data objects and flows. *Big Data & Society*, *3*(2), 2053951716654502. https://doi.org/10.1177/2053951716654502

Bauer, M. W. (2000). Classical content analysis: A review. *Qualitative Researching with Text, Image and Sound*, 131–151.

Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Archives of Disease in Childhood-Education and Practice*, *98*(6), 236–238.

BenderlySep. 5, B. L., 2008, & Am, 8:00. (2008, September 5). *Fitting the Job Market to a T*. Science | AAAS. https://www.sciencemag.org/careers/2008/09/fitting-job-market-t

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2017). GenBank. *Nucleic Acids Research*, *45*(D1), D37–D42. https://doi.org/10.1093/nar/gkw1070

Berg, B. L., & Lune, H. (2012). *Qualitative Research Methods for the Social Sciences. 8 utg*. Boston: Pearson.

Bietz, M. J., & Lee, C. P. (2009). Collaboration in metagenomics: Sequence databases and the organization of scientific work. In *ECSCW 2009* (pp. 243–262). Springer.

Bishop, B. W., & Hank, C. (2018). Measuring FAIR Principles to Inform Fitness for Use. *International Journal of Digital Curation*, *13*(1), 35–46. https://doi.org/10.2218/ijdc.v13i1.630

Bishop, B. W., Ungvari, J., Gunderman, H., & Moulaison-Sandy, H. (2020). Data management plan scorecard. *Proceedings of the Association for Information Science and Technology*, *57*(1), e325.

Bloor, D. (1984). The strengths of the strong programme. In *Scientific rationality: The sociological turn* (pp. 75–94). Springer.

Bobko, P., Barelka, A., & Hirshfield, L. M. (2014). The construct of state-level suspicion: A model and research agenda for automated and information technology (IT) contexts. *Human Factors*, *56*(3), 489-508.

Boden, A., Nett, B., & Wulf, V. (2008). Articulation work in small-scale offshore software development projects. *Proceedings of the 2008 International Workshop on Cooperative and Human Aspects of Software Engineering - CHASE '08*, 21–24. https://doi.org/10.1145/1370114.1370120

Borgman, C. L. (1990). *Scholarly communication and bibliometrics*. Sage Publications.

Borgman, C. L. (2000). Digital libraries and the continuum of scholarly communication. *Journal of Documentation*, *56*(4), 412–430.

Borgman, C. L. (2007). *Scholarship in the digital age: Information, infrastructure, and the Internet*. MIT press.

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. MIT press.

Borgman, C. L. (2019). The lives and after lives of data. *Harvard Data Science Review*, *1*(1).

Borgman, C. L., Darch, P. T., Sands, A. E., Pasquetto, I. V., Golshan, M. S., Wallis, J. C., & Traweek, S. (2015). Knowledge infrastructures in science: Data, diversity, and digital libraries. *International Journal on Digital Libraries*, *16*(3–4), 207–227.

Borgman, C. L., Scharnhorst, A., & Golshan, M. S. (2019). Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. *Journal of the Association for Information Science and Technology*, *70*(8), 888–904. https://doi.org/10.1002/asi.24172

Borgman, C. L., Wallis, J. C., & Mayernik, M. S. (2012). Who's got the data? Interdependencies in science and technology collaborations. *Computer Supported Cooperative Work (CSCW)*, *21*(6), 485–523.

Bossen, C., Chen, Y., & Pine, K. H. (2019). The emergence of new data work occupations in healthcare: The case of medical scribes. *International Journal of Medical Informatics*, *123*, 76–83.

Bowker, G. C. (2000). Biodiversity datadiversity. *Social Studies of Science*, *30*(5), 643–683.

Bowker, G. C. (2005). *Memory practices in the sciences*. Mit Press Cambridge, MA.

Bowker, G. C., & Star, S. L. (2000). *Sorting things out: Classification and its consequences*. MIT press. https://books.google.com/books?hl=en&lr=&id=xHlP8WqzizYC&oi=fnd&pg=PR9&dq=bowker+and+star&ots=MA0zpzlYsJ&sig=windJpbyBNqGDeFi_F_v9WirZsM

Bowker, G. C., Star, S. L., & Spasser, M. (2001). Classifying nursing work. *Online Journal of Issues in Nursing*, *6*(2).

Bowker, G. C., Timmermans, S., Clarke, A. E., & Balka, E. (2016). *Boundary objects and beyond: Working with Leigh Star*. MIT Press.

Boyer, E. L. (1997). Scholarship reconsidered: Priorities of the Professoriate. 1990. *Princeton, NJ: Carnegie Foundation for the Advancement of Teaching*.

Bozeman, B., & Boardman, C. (2014a). Assessing research collaboration studies: A framework for analysis. In *Research collaboration and team science* (pp. 1–11). Springer.

Bozeman, B., & Boardman, C. (2014b). *Research collaboration and team science: A state-of-the-art review and agenda*. Springer.

Bratt, S., Hemsley, J., Qin, J., & Costa, M. (2017). Big data, big metadata and quantitative study of science: A workflow model for big scientometrics. *Proceedings of the Association for Information Science and Technology*, *54*(1), 36–45.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101.

Buckland, M. K. (1991). Information as thing. *Journal of the American Society for Information Science (1986-1998)*, *42*(5), 351.

Byrd, J. B., Greene, A. C., Prasad, D. V., Jiang, X., & Greene, C. S. (2020). Responsible, practical genomic data sharing that accelerates research. *Nature Reviews Genetics*, *21*(10), 615–629.

Calude, C. S., & Longo, G. (2017). The deluge of spurious correlations in big data. *Foundations of Science*, *22*(3), 595–612.

Campbell, P. (2009). Data's shameful neglect. *Nature*, *461*(7261), 145.

Campos-Mercade, P., Meier, A. N., Schneider, F. H., & Wengström, E. (2021). Prosociality predicts health behaviors during the COVID-19 pandemic. *Journal of Public Economics*, *195*, 104367.

Carlson, J. (2014). The use of life cycle models in developing and supporting data services. *Research Data Management: Practical Strategies for Information Professionals*, 63–86.

Carroll, S. R., Herczog, E., Hudson, M., Russell, K., & Stall, S. (2021). Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Scientific Data*, *8*(1), 108. https://doi.org/10.1038/s41597-021-00892-0

Casneuf, T., Van de Peer, Y., & Huber, W. (2007). In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. *BMC Bioinformatics*, *8*(1), 461.

Ceci, S. J. (1988). Scientists' attitudes toward data sharing. *Science, Technology, & Human Values*, *13*(1–2), 45–52.

Charmaz, K. (2006). Constructing grounded theory: A practical guide through qualitative research. *SagePublications Ltd, London*.

Check Hayden, E. (n.d.). Funding for model-organism databases in trouble. *Nature News*. https://doi.org/10.1038/nature.2016.20134

Chen, W., Liang, X., Li, J., Qin, H., Mu, Y., & Wang, J. (2018). Blockchain Based Provenance Sharing of Scientific Workflows. *2018 IEEE International Conference on Big Data (Big Data)*, 3814–3820. https://doi.org/10.1109/BigData.2018.8622237

Church, S. P., Dunn, M., & Prokopy, L. S. (2019). Benefits to qualitative data quality with multiple coders: Two case studies in multi-coder data analysis. *Journal of Rural Social Sciences*, *34*(1), 2.

Ciborra, C. U. (1992). From thinking to tinkering: The grassroots of strategic information systems. *The Information Society*, *8*(4), 297–309.

Cole, S. (2004). Merton's Contribution to the Sociology of Science. *Social Studies of Science*, *34*(6), 829–844.

Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: Lessons from large-scale biology. *Science*, *300*(5617), 286–290.

Collins, H., & Evans, R. (2008). *Rethinking expertise*. University of Chicago Press.

Collins, H., & Pinch, T. (2002). *The Golem at Large: What You Should Know about Technology* (Reprint edition). Cambridge University Press.

Corbin, J. M., & Strauss, A. L. (1993). The articulation of work through interaction. *The Sociological Quarterly*, *34*(1), 71–83.

Corpas, M., Kovalevskaya, N. V., McMurray, A., & Nielsen, F. G. G. (2018). A FAIR guide for data providers to maximise sharing of human genomic data. *PLOS Computational Biology*, *14*(3), e1005873. https://doi.org/10.1371/journal.pcbi.1005873

Costa, M. R., Qin, J., & Bratt, S. (2016). Emergence of collaboration networks around large scale data repositories: A study of the genomics community using GenBank. *Scientometrics*, *108*(1), 21–40. https://doi.org/10.1007/s11192-016-1954-x

Cox, A. M., & Tam, W. W. T. (2018). A critical analysis of lifecycle models of the research process and research data management. *Aslib Journal of Information Management*, *70*(2), 142–157.

Cox, J. (2016). Communicating new library roles to enable digital scholarship: A review article. *New Review of Academic Librarianship*, *22*(2–3), 132–147.

Creswell, J. W., Hanson, W. E., Clark Plano, V. L., & Morales, A. (2007). Qualitative research designs: Selection and implementation. *The Counseling Psychologist*, *35*(2), 236–264.

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

Cronin, B., & Sugimoto, C. R. (2014). *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*. MIT Press.

Crowston, K., & Qin, J. (2011). A capability maturity model for scientific data management: Evidence from the literature. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1–9. https://doi.org/10.1002/meet.2011.14504801036

Cui, H., Macklin, J. A., Sachs, J., Reznicek, A., Starr, J., Ford, B., Penev, L., & Chen, H.-L. (2018). Incentivising use of structured language in biological descriptions: Author-driven

phenotype data and ontology production. *Biodiversity Data Journal*, *6*, e29616. https://doi.org/10.3897/BDJ.6.e29616

Curty, R. G. (2015). *Beyond "Data Thrifting": An Investigation of Factors Influencing Research Data Reuse In the Social Sciences*.

Curty, R., Kim, Y., & Qin, J. (2013). *What have scientists planned for data sharing and reuse? A content analysis of NSF awardees' data management plans*.

Cyranoski, D. (2021). Alarming COVID variants show vital role of genomic surveillance. *Nature*, *589*(7842), 337–338. https://doi.org/10.1038/d41586-021-00065-4

Daniels, K., Johnson, G., & De Chernatony, L. (2002). Task and institutional influences on managers' mental models of competition. *Organization Studies*, *23*(1), 31–62.

Darch, P. T., Borgman, C. L., Traweek, S., Cummings, R. L., Wallis, J. C., & Sands, A. E. (2015). What lies beneath?: Knowledge infrastructures in the subseafloor biosphere and beyond. *International Journal on Digital Libraries*, *16*(1), 61–77. https://doi.org/10.1007/s00799-015-0137-3

Darch, P. T., Sands, A. E., Borgman, C. L., & Golshan, M. S. (2020a). Do the stars align?: Stakeholders and strategies in libraries' curation of an astronomy dataset. *Journal of the Association for Information Science and Technology*.

Darch, P. T., Sands, A. E., Borgman, C. L., & Golshan, M. S. (2020b). Library Cultures of Data Curation: Adventures in Astronomy. *Journal of the Association for Information Science and Technology*.

Darch, P. T., Sands, A. E., Borgman, C. L., & Golshan, M. S. (2020c). Library Cultures of Data Curation: Adventures in Astronomy. *Journal of the Association for Information Science and Technology*.

Data sharing and the future of science. (2018). *Nature Communications*, *9*(1), 2817. https://doi.org/10.1038/s41467-018-05227-z

Daugelaite, J., O'Driscoll, A., & Sleator, R. D. (2013). An overview of multiple sequence alignments and cloud computing in bioinformatics. *International Scholarly Research Notices*, *2013*.

Deleuze, G., & Guattari, F. (1988). *A thousand plateaus: Capitalism and schizophrenia*. Bloomsbury Publishing.

Demchenko, Y., & Stoy, L. (2021). Research Data Management and Data Stewardship Competences in University Curriculum. *2021 IEEE Global Engineering Education Conference (EDUCON)*, 1717–1726.

Den Besten, M., David, P. A., & Schroeder, R. (2009). Research in e-science and open access to data and information. In *International Handbook of Internet Research* (pp. 65–96). Springer.

Denis, J., Mongili, A., & Pontille, D. (2016). Maintenance & Repair in Science and Technology Studies. *TECNOSCIENZA: Italian Journal of Science & Technology Studies*, *6*(2), 5-16–16.

Diekema, A. R., Wesolek, A., & Walters, C. D. (2014). The NSF/NIH Effect: Surveying the Effect of Data Management Requirements on Faculty, Sponsored Programs, and Institutional Repositories. *The Journal of Academic Librarianship*, *40*(3), 322–331. https://doi.org/10.1016/j.acalib.2014.04.010

DiMaggio, P. J. (1991). Introduction. I: PJ DiMaggio & W. W. Powell. *The New Institutionalism in Organizational Analysis*.

DiMaggio, P. J., & Powell, W. W. (2000). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields (American Sociological Review, 1983). *Advances in Strategic Management*, *17*, 143–166.

Douglas, M. (1986). *How institutions think*. Syracuse University Press.

Dowling, J., & Pfeffer, J. (1975). Organizational legitimacy: Social values and organizational behavior. *Pacific Sociological Review*, *18*(1), 122–136.

Durmaz, A. A., Karaca, E., Demkow, U., Toruner, G., Schoumans, J., & Cogulu, O. (2015, March 22). *Evolution of Genetic Techniques: Past, Present, and Beyond* [Review Article]. BioMed Research International; Hindawi. https://doi.org/10.1155/2015/461524

Duxbury, L., & Haines Jr, G. (1991). Predicting alternative work arrangements from salient attitudes: A study of decision makers in the public sector. *Journal of Business Research*, *23*(1), 83–97.

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. Mit Press.

Edwards, P. N. (2017). Knowledge infrastructures for the Anthropocene. *The Anthropocene Review*, *4*(1), 34–43. https://doi.org/10.1177/2053019616679854

Edwards, P. N., Jackson, S. J., Bowker, G. C., & Knobel, C. P. (2007). *Understanding infrastructure: Dynamics, tensions, and design*.

Elo, S., & Kyngäs, H. (2008a). The qualitative content analysis process. *Journal of Advanced Nursing*, *62*(1), 107–115.

Elo, S., & Kyngäs, H. (2008b). The qualitative content analysis process. *Journal of Advanced Nursing*, *62*(1), 107–115.

Elragal, A., & Klischewski, R. (2017). Theory-driven or process-driven prediction? Epistemological challenges of big data analytics. *Journal of Big Data*, *4*(1), 1–20.

Engeström, Y. (1990). *When is a tool?*

Erickson, I., & Sawyer, S. B. (2019). Infrastructuring as bricolage: Thinking like a contemporary knowledge worker. In *Research in the Sociology of Organizations* (pp. 321–334). Emerald Group Publishing Ltd. https://doi.org/10.1108/S0733-558X20190000062020

Fehr, A. von der, Sølberg, J., & Bruun, J. (2018). Validation of networks derived from snowball sampling of municipal science education actors. *International Journal of Research & Method in Education*, *41*(1), 38–52. https://doi.org/10.1080/1743727X.2016.1192117

Feldman, M. S., & Pentland, B. T. (2003). Reconceptualizing organizational routines as a source of flexibility and change. *Administrative Science Quarterly*, *48*(1), 94–118.

Feldman, M. S., Pentland, B. T., D'Adderio, L., & Lazaric, N. (2016). *Beyond routines as things: Introduction to the special issue on routine dynamics*. INFORMS.

Felt, U., Fouché, R., Miller, C. A., & Smith-Doerr, L. (2017). *The handbook of science and technology studies*. Mit Press.

Fiesler, C., Beard, N., & Keegan, B. C. (2020). No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 187–196.

Freese, J., & King, M. M. (2018). Institutionalizing Transparency. *Socius*, *4*, 2378023117739216. https://doi.org/10.1177/2378023117739216

Friedland, R. (1991). Bringing society back in: Symbols, practices, and institutional contradictions. *The New Institutionalism in Organizational Analysis*, 232–263.

Fujimura, J. H. (1987). Constructingdo-Able'problems in Cancer research: Articulating alignment. *Social Studies of Science*, *17*(2), 257–293.

Furner, J. (2004). Information studies without information. *Library Trends*, *52*(3), 427.

Garnett, F., & Ecclesfield, N. (2011). Towards a framework for co-creating open scholarship. *Research in Learning Technology*, *19*.

Geiger, R. S., Sholler, D., Culich, A., Martinez, C., Guardia, F. H. de la, Lanusse, F., Ottoboni, K., Stuart, M., Vareth, M., & Varoquaux, N. (2018). *Challenges of Doing Data-Intensive Research in Teams, Labs, and Groups: Report from the BIDS Best Practices in Data Science Series*. https://doi.org/10.17605/OSF.IO/UV6FY

George, E., Chattopadhyay, P., Sitkin, S. B., & Barden, J. (2006). Cognitive underpinnings of institutional persistence and change: A framing perspective. *Academy of Management Review*, *31*(2), 347–365.

Gerson, E. M. (1983). Scientific work and social worlds. *Knowledge*, *4*(3), 357–377.

Gerson, E. M., & Star, S. L. (1986). Analyzing due process in the workplace. *ACM Transactions on Information Systems (TOIS)*, *4*(3), 257–270.

Gile, K. J., & Handcock, M. S. (2010). 7. Respondent-driven sampling: An assessment of current methodology. *Sociological Methodology*, *40*(1), 285–327.

Glaser, B. G. (1965). The constant comparative method of qualitative analysis. *Social Problems*, *12*(4), 436–445.

Glaser, B. G., & Strauss, A. L. (2017). *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

Gold, A. K. (2007). Cyberinfrastructure, data, and libraries, part 1: A cyberinfrastructure primer for librarians. *Office of the Dean (Library)*, 16.

Goldsmith, M. (1967). The autonomy of science: Some thoughts for discussion. *The Political Quarterly*, *38*(1), 81–89.

Gonzales, J. A., Fiesler, C., & Bruckman, A. (2015). Towards an Appropriable CSCW Tool Ecology: Lessons from the Greatest International Scavenger Hunt the World Has Ever Seen. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 946–957. https://doi.org/10.1145/2675133.2675240

Gray, J. (2009). Jim Gray on eScience: A transformed scientific method. *The Fourth Paradigm: Data-Intensive Scientific Discovery*, *1*.

Gray, M. L., & Suri, S. (2019). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Eamon Dolan Books.

Greenberg, J. (2009). Theoretical Considerations of Lifecycle Modeling: An Analysis of the Dryad Repository Demonstrating Automatic Metadata Propagation, Inheritance, and Value System Adoption. *Cataloging & Classification Quarterly*, *47*(3–4), 380–402. https://doi.org/10.1080/01639370902737547

Greenwood, R., Oliver, C., Lawrence, T. B., & Meyer, R. E. (2017). *The Sage handbook of organizational institutionalism*. Sage.

Griffin, P. C., Khadake, J., LeMay, K. S., Lewis, S. E., Orchard, S., Pask, A., Pope, B., Roessner, U., Russell, K., Seemann, T., Treloar, A., Tyagi, S., Christiansen, J. H., Dayalan, S., Gladman, S., Hangartner, S. B., Hayden, H. L., Ho, W. W. H., Keeble-Gagnère, G., … Schneider, M. V. (2018). Best practice data life cycle approaches for the life sciences. *F1000Research*, *6*. https://doi.org/10.12688/f1000research.12344.2

Gupta, A., Lai, A., Mozersky, J., Ma, X., Walsh, H., & DuBois, J. M. (2021). Enabling qualitative research data sharing using a natural language processing pipeline for

deidentification: Moving beyond HIPAA Safe Harbor identifiers. *JAMIA Open*, *4*(3), ooab069.

Hall, K. L., Vogel, A. L., Huang, G. C., Serrano, K. J., Rice, E. L., Tsakraklides, S. P., & Fiore, S. M. (2018). The science of team science: A review of the empirical evidence and research gaps on collaboration in science. *American Psychologist*, *73*(4), 532–548. https://doi.org/10.1037/amp0000319

Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M. T., & Beyene, J. (2009). Data Integration in Genetics and Genomics: Methods and Challenges. *Human Genomics and Proteomics : HGP*, *2009*. https://doi.org/10.4061/2009/869093

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381.

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends*, *57*(2), 280–299.

Hemsley, J., Qin, J., & Bratt, S. E. (2020). Data to knowledge in action: A longitudinal analysis of GenBank metadata. *Proceedings of the Association for Information Science and Technology*, *57*(1), e253.

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. Springer Science & Business Media.

Hesse, B. W., Croyle, R. T., & Buetow, K. H. (2011). Cyberinfrastructure and the Biomedical Sciences. *American Journal of Preventive Medicine*, *40*(5), S97–S102. https://doi.org/10.1016/j.amepre.2011.01.006

Heugens, P. P., & Lander, M. W. (2009). Structure! Agency!(and other quarrels): A meta-analysis of institutional theories of organization. *Academy of Management Journal*, *52*(1), 61–85.

Hey, T., Tansley, S., & Tolle, K. M. (2009). *The fourth paradigm: Data-intensive scientific discovery* (Vol. 1). Microsoft research Redmond, WA.

Hey, T., & Trefethen, A. E. (2005). Cyberinfrastructure for e-Science. *Science*, *308*(5723), 817–821.

Hine, C. (2005). Virtual methods and the sociology of cyber-social-scientific knowledge. *Virtual Methods: Issues in Social Research on the Internet*, 1–13.

Hine, C. (2006). *New infrastructures for knowledge production: Understanding e-science*. IGI Global.

Hood, L., & Rowen, L. (2013). The Human Genome Project: Big science transforms biology and medicine. *Genome Medicine*, *5*(9), 79. https://doi.org/10.1186/gm483

Hrynaszkiewicz, I., Simons, N., Hussain, A., Grant, R., & Goudie, S. (2020). Developing a research data policy framework for all journals and publishers. *Data Science Journal*, *19*(1).

Humphrey, W. S. (1989). *Managing the software process*. Addison-Wesley Longman Publishing Co., Inc.

Illenberger, J., & Flötteröd, G. (2012). Estimating network properties from snowball sampled data. *Social Networks*, *34*(4), 701–711.

Jeng, W., He, D., & Chi, Y. (2017). Social science data repositories in data deluge: A case study of ICPSR's workflow and practices. *The Electronic Library*.

Jennex, M. E. (2009). Re-visiting the knowledge pyramid. *2009 42nd Hawaii International Conference on System Sciences*, 1–7.

Jiang, J. A., Wade, K., Fiesler, C., & Brubaker, J. R. (2021). Supporting Serendipity: Opportunities and Challenges for Human-AI Collaboration in Qualitative Analysis. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–23.

Jones, K. M., Ankeny, R. A., & Cook-Deegan, R. (2018). The Bermuda Triangle: The pragmatics, policies, and principles for data sharing in the history of the Human Genome Project. *Journal of the History of Biology*, *51*(4), 693.

Joo, S., & Peters, C. (2020). User needs assessment for research data services in a research university. *Journal of Librarianship and Information Science*, *52*(3), 633–646.

Kanza, S., Willoughby, C., Gibbins, N., Whitby, R., Frey, J. G., Erjavec, J., Zupančič, K., Hren, M., & Kovač, K. (2017). Electronic lab notebooks: Can they replace paper? *Journal of Cheminformatics*, *9*(1), 31. https://doi.org/10.1186/s13321-017-0221-3

Katz, J. S. (1994). Geographical proximity and scientific collaboration. *Scientometrics*, *31*(1), 31–43.

Kelemen, Z. D., Kusters, R., Trienekens, J., & Balla, K. (2013). *Towards complexity analysis of software process improvement frameworks*. Budapest, Technical Report TR201301.

Kelle, U. (2005). “Emergence” vs. “Forcing” of Empirical Data? A Crucial Problem of “Grounded Theory” Reconsidered. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *6*(2), Article 2. https://doi.org/10.17169/fqs-6.2.467

Kellogg, K. C., Orlikowski, W. J., & Yates, J. (2006). Life in the trading zone: Structuring coordination across boundaries in postbureaucratic organizations. *Organization Science*, *17*(1), 22–44.

Keman, H. (2017). Institutionalization. *Encyclopedia Britannica Retrieved from, Https://Www. Britannica. Com/Topic/Institutionalization*.

Khan, N., Pink, C. J., & Thelwall, M. (2020). Identifying Data Sharing and Reuse with Scholix: Potentials and Limitations. *Patterns*, *1*(1), 100007.

Kim, Y. (2013). Institutional and Individual Influences on Scientists' Data Sharing Behaviors. *School of Information Studies - Dissertations*. https://surface.syr.edu/it_etd/85

Kim, Y., & 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. https://doi.org/10.1016/j.ijinfomgt.2015.04.007

Kim, Y., & Stanton, J. M. (2016). Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology*, *67*(4), 776–799. https://doi.org/10.1002/asi.23424

Kitamoto, A. (2017). Digital Typhoon and open science—A trans-disciplinary platform for typhoon-related data. *Japan Geoscience Union Meeting*.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 2053951714528481. https://doi.org/10.1177/2053951714528481

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J. B., Grout, J., & Corlay, S. (2016). Jupyter Notebooks-a publishing format for reproducible computational workflows. *ELPUB*, 87–90.

Kolb, S. M. (2012). Grounded theory and the constant comparative method: Valid research strategies for educators. *Journal of Emerging Trends in Educational Research and Policy Studies*, *3*(1), 83–86.

Kollen, C., Kouper, I., Ishida, M., Williams, S., & Fear, K. (2017). *Research Data Services Maturity in Academic Libraries*. American Library Association, Association of College and Research Libraries. https://repository.arizona.edu/handle/10150/622168

Koulikoff-Souviron, M., & Harrison, A. (2008). Interdependent supply relationships as institutions: The role of HR practices. *International Journal of Operations & Production Management*, *28*(5), 412–432. https://doi.org/10.1108/01443570810867187

Kowalczyk, S., & Shankar, K. (2011). Data sharing in the sciences. *Annual Review of Information Science and Technology*, *45*(1), 247–294. https://doi.org/10.1002/aris.2011.1440450113

Kowalczyk, S. T. (2018). Modelling the Research Data Lifecycle. *International Journal of Digital Curation*, *12*(2), 331–361. https://doi.org/10.2218/ijdc.v12i2.429

Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor Protocols*, *2015*(11), 951–969. https://doi.org/10.1101/pdb.top084970

Lai, ronald, D'Amour, A., Yu, A., & Fleming, L. (2011). *Disambiguation and Co-authorship Networks of the U.S. Patent Inventor Database (1975—2010)* [Dataset]. Patent Network Dataverse (Harvard Business School). https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/15705

Lakhani, K. R., Lifshitz-Assaf, H., & Tushman, M. L. (2013). Open innovation and organizational boundaries: Task decomposition, knowledge distribution and the locus of innovation. In *Handbook of economic organization*. Edward Elgar Publishing.

Larivière, V., Desrochers, N., Macaluso, B., Mongeon, P., Paul-Hus, A., & Sugimoto, C. R. (2016). Contributorship and division of labor in knowledge production. *Social Studies of Science*, *46*(3), 417–435. https://doi.org/10.1177/0306312716650046

Larsen, R. L., Palmer, C., Lyon, L., Hedstrom, M., & de Roure, D. (2014). Preparing the workforce for digital curation. *Panel Presented at the 9th International Digital Curation Conference, San Francisco*.

Latour, B. (1987). *Science in Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press.

Latour, B., & Woolgar, S. (1986). *Laboratory Life: The Construction of Scientific Facts*.

Latour, B., & Woolgar, S. (2013). *Laboratory life: The construction of scientific facts*. Princeton University Press.

Lawrence, T. B., & Suddaby, R. (2006). 1.6 institutions and institutional work. *The Sage Handbook of Organization Studies*, 215–254.

Lawrence, T. B., Winn, M. I., & Jennings, P. D. (2001). The temporal dynamics of institutionalization. *Academy of Management Review*, *26*(4), 624–644.

Lee, C. A. (2010). Open archival information system (OAIS) reference model. *Encyclopedia of Library and Information Sciences, 3*.

Lee, C. P., Dourish, P., & Mark, G. (2006). The human infrastructure of cyberinfrastructure. *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, 483–492. http://dl.acm.org/citation.cfm?id=1180950

Leonelli, S. (2010). *Packaging Data for Re-Use: Databases in Model Organism Biology*. Cambridge University Press.

Leonelli, S. (2014a). What Difference Does Quantity Make? On the Epistemology of Big Data in Biology. *Big Data & Society*, *1*(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4340542/

Leonelli, S. (2014b). What Difference Does Quantity Make? On the Epistemology of Big Data in Biology. *Big Data & Society*, *1*(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4340542/

Leonelli, S. (2016). *Data-centric biology: A philosophical study*. University of Chicago Press.

Leray, M., Knowlton, N., Ho, S.-L., Nguyen, B. N., & Machida, R. J. (2019). GenBank is a reliable resource for 21st century biodiversity research. *Proceedings of the National Academy of Sciences*, *116*(45), 22651–22656. https://doi.org/10.1073/pnas.1911714116

LEVINE, M. (2014). Copyright, Open Data, and the Availability-Usability Gap. *Research Data Management: Practical Strategies for Information Professionals*, 129.

Louis, K. S., Jones, L. M., & Campbell, E. G. (2002). Macroscope: Sharing in science. *American Scientist*, *90*(4), 304–307.

Lounsbury, M. (2001). Institutional sources of practice variation: Staffing college and university recycling programs. *Administrative Science Quarterly*, *46*(1), 29–56.

Lounsbury, M., Steele, C. W., Wang, M. S., & Toubiana, M. (2021). New Directions in the Study of Institutional Logics: From Tools to Phenomena. *Annual Review of Sociology*, *47*.

Luo, X. (2007). Continuous learning: The influence of national institutional logics on training attitudes. *Organization Science*, *18*(2), 280–296.

Mannheimer, S., Pienta, A., Kirilova, D., Elman, C., & Wutich, A. (2019). Qualitative data sharing: Data repositories and academic libraries as key partners in addressing challenges. *American Behavioral Scientist*, *63*(5), 643–664.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, *24*(3), 133–141.

Marquis, C., & Tilcsik, A. (2013). Imprinting: Toward a multilevel theory. *Academy of Management Annals*, *7*(1), 195–245.

Marshall, E. (2001). *Bermuda rules: Community spirit, with teeth*. American Association for the Advancement of Science.

Mayernik, M. S., Batcheller, A. L., & Borgman, C. L. (2011). How institutional factors influence the creation of scientific metadata. *Proceedings of the 2011 IConference*, 417–425. http://dl.acm.org/citation.cfm?id=1940818

McCain, K. W. (1991). Communication, competition, and secrecy: The production and dissemination of research-related information in genetics. *Science, Technology, & Human Values*, *16*(4), 491–516.

McDonald, N., Schoenebeck, S., & Forte, A. (2019). Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–23. https://doi.org/10.1145/3359174

McGrath, P. J. (2002). *Scientists, business, and the state, 1890-1960*. Univ of North Carolina Press.

Meadows, A. J. (1997). *Communicating research*. Emerald Group Publishing Limited.

Mellor, D. (2017). Promoting reproducibility with registered reports. *Nature Human Behaviour*, *1*.

Meng, X.-L. (2019). Data Science: An Artificial Ecosystem. *1.1*, *1*(1). https://doi.org/10.1162/99608f92.ba20f892

Merrill, S. A., Mazza, A.-M., & Innovation, N. R. C. (US) C. on I. P. R. in G. and P. R. and. (2006). *Genomics, Proteomics, and the Changing Research Environment*. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK19861/

Merton, R. K. (1957). Priorities in scientific discovery: A chapter in the sociology of science. *American Sociological Review*, *22*(6), 635–659.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago press.

Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, *83*(2), 340–363.

Michels, D. (n.d.). *Going Pi-Shaped: How To Prepare For The Work Of The Future*. Forbes. Retrieved November 28, 2020, from https://www.forbes.com/sites/davidmichels/2019/09/27/going-pi-shaped-how-to-prepare-for-the-work-of-the-future/

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: Supporting ecology as a data-intensive science. *Trends in Ecology & Evolution*, *27*(2), 85–93. https://doi.org/10.1016/j.tree.2011.11.016

Mirowski, P. (2018). The future(s) of open science. *Social Studies of Science*, *48*(2), 171–203. https://doi.org/10.1177/0306312718772086

Mitroff, I. I. (1974). Norms and counter-norms in a select group of the Apollo moon scientists: A case study of the ambivalence of scientists. *American Sociological Review*, 579–595.

Mozersky, J., McIntosh, T., Walsh, H. A., Parsons, M. V., Goodman, M., & DuBois, J. M. (2021). Barriers and facilitators to qualitative data sharing in the United States: A survey of qualitative researchers. *Plos One*, *16*(12), e0261719.

Murray, F., & O'Mahony, S. (2007). Exploring the foundations of cumulative innovation: Implications for organization science. *Organization Science*, *18*(6), 1006–1021.

Myers, K., Tham, W. Y., Yin, Y., Cohodes, N., Thursby, J. G., Thursby, M., Schiffer, P., Walsh, J., Lakhani, K. R., & Wang, D. (2020). Quantifying the Immediate Effects of the COVID-19 Pandemic on Scientists. *Available at SSRN 3608302*.

Nadim, T. (2016). Data labours: How the sequence databases GenBank and EMBL-Bank make data. *Science as Culture*, *25*(4), 496–519.

Nardi, B. A., & O'Day, V. (1999). *Information ecologies: Using technology with heart*. Mit Press.

Navale, V., & McAuliffe, M. (2018). Long-term preservation of biomedical research data. *F1000Research*, *7*, 1353. https://doi.org/10.12688/f1000research.16015.1

Neang, A., Sutherland, W., Beach, M., & Lee, C. (2020). *Data Integration as Coordination: The Articulation of Data Work in an Ocean Science Collaboration*.

Ni, C., Sugimoto, C. R., & Cronin, B. (2013). Visualizing and comparing four facets of scholarly communication: Producers, artifacts, concepts, and gatekeepers. *Scientometrics*, *94*(3), 1161–1173. https://doi.org/10.1007/s11192-012-0849-8

North, D. C. (1991). Institutions. *Journal of Economic Perspectives*, *5*(1), 97–112.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606.

Now, C. (2016, March 3). GenBank & The Early Years of "Big Data." *Circulating Now from NLM*. https://circulatingnow.nlm.nih.gov/2016/03/03/genbank-the-early-years-of-big-data/

Nowakowska, J., Sobocińska, J., Lewicki, M., Lemańska, Ż., & Rzymski, P. (2020). When science goes viral: The research response during three months of the COVID-19 outbreak. *Biomedicine & Pharmacotherapy*, *129*, 110451. https://doi.org/10.1016/j.biopha.2020.110451

Ocasio, W. (1994). Political dynamics and the circulation of power: CEO succession in US industrial corporations, 1960-1990. *Administrative Science Quarterly*, 285–312.

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*, *19*, 1609406919899220.

O'Hara, R. J. (1992). Telling the tree: Narrative representation and the study of evolutionary history. *Biology and Philosophy*, *7*(2), 135–160.

Oliver, P., & Jupp, V. (2006). *Purposive sampling*. Sage.

Oliver, S. G., Lock, A., Harris, M. A., Nurse, P., & Wood, V. (2016). Model organism databases: Essential resources that need the support of both funders and users. *BMC Biology*, *14*(1), 49. https://doi.org/10.1186/s12915-016-0276-z

On the Big Impact of "Big Computer Science" (pp. 17–26). (2017). https://doi.org/10.1007/978-3-319-55735-9_2

Østerlund, C., & Carlile, P. (2005). Relations in Practice: Sorting Through Practice Theories on Knowledge Sharing in Complex Organizations. *The Information Society*, *21*(2), 91–107. https://doi.org/10.1080/01972240590925294

Paisley, W. J. (1968). *As We May Think, Information Systems Do Not.*

Paulk, M. C. (2008). A taxonomy for improvement frameworks. *Fourth World Congress for Software Quality*, 15–18.

Paulk, M. C. (2009). A history of the capability maturity model for software. *ASQ Software Quality Professional*, *12*(1), 5–19.

Paulk, M. C., Curtis, B., Chrissis, M. B., & Weber, C. V. (1993). Capability maturity model, version 1.1. *IEEE Software*, *10*(4), 18–27.

Pickard, A. J. (2013). *Research methods in information*. Facet publishing.

Pinel, C., Prainsack, B., & McKevitt, C. (2020). Caring for data: Value creation in a data-intensive research laboratory. *Social Studies of Science*, *50*(2), 175–197. https://doi.org/10.1177/0306312720906567

Plantin, J.-C. (2019). Data cleaners for pristine datasets: Visibility and invisibility of data processors in social science. *Science, Technology, & Human Values*, *44*(1), 52–73.

Polanyi, M. (1945). The autonomy of science. *The Scientific Monthly*, *60*(2), 141–150.

Porter, M. E. (1985). Technology and competitive advantage. *The Journal of Business Strategy*, *5*(3), 60.

Price, D. de S. (1963). Big science, little science. *Columbia University, New York*, 119–119.

Priem, J. (2014). Altmetrics. *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*, 263–288.

Pujol Priego, L., Wareham, J., & Romasanta, A. K. S. (2022). The puzzle of sharing scientific data. *Industry and Innovation*, *29*(2), 219–250. https://doi.org/10.1080/13662716.2022.2033178

Qin, J. (2013). *Infrastructure, standards, and policies for research data management*.

Qin, J., Costa, M., & Wang, J. (2015). Methodological and Technical Challenges in Big Scientometric Data Analytics. *IConference 2015 Proceedings*. https://www.ideals.illinois.edu/handle/2142/73756

Raffaghelli, J. E. (2017). Exploring the (missed) connections between digital scholarship and faculty development: A conceptual analysis. *International Journal of Educational Technology in Higher Education*, *14*(1), 20. https://doi.org/10.1186/s41239-017-0058-x

Raffaghelli, J. E., Cucchiara, S., Manganello, F., & Persico, D. (2016). Different views on Digital Scholarship: Separate worlds or cohesive research field? *Research in Learning Technology*, *24*.

Randles, B. M., Pasquetto, I. V., Golshan, M. S., & Borgman, C. L. (2017). Using the Jupyter Notebook as a Tool for Open Science: An Empirical Study. *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 1–2. https://doi.org/10.1109/JCDL.2017.7991618

Ray, J. M. (2013). *Research data management: Practical strategies for information professionals*. Purdue University Press.

Read, K. B., Larson, C., Gillespie, C., Oh, S. Y., & Surkis, A. (2019). A two-tiered curriculum to improve data management practices for researchers. *PloS One*, *14*(5), e0215509.

Rehm, G., Bourgonje, P., Hegele, S., Kintzel, F., Schneider, J. M., Ostendorff, M., Zaczynska, K., Berger, A., Grill, S., Räuchle, S., Rauenbusch, J., Rutenburg, L., Schmidt, A., Wild, M., Hoffmann, H., Fink, J., Schulz, S., Seva, J., Quantz, J., … Heine, F. (2020). QURATOR: Innovative Technologies for Content and Data Curation. *ArXiv:2004.12195 [Cs]*. http://arxiv.org/abs/2004.12195

Renear, A. H., Sacchi, S., & Wickett, K. M. (2010). Definitions of dataset in the scientific and technical literature. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–4.

Ribes, D. (2019). STS, meet data science, once again. *Science, Technology, & Human Values*, *44*(3), 514–539.

Ribes, D., & Bowker, G. C. (2008). *Organizing for multidisciplinary collaboration: The case of the geosciences network*.

Ribes, D., & Finholt, T. A. (2009). *The long now of infrastructure: Articulating tensions in development*.

Rivera, M. A. (2020). *Big data research in hospitality: From streetlight empiricism research to theory laden research*. Elsevier.

Rougier, N. P., Hinsen, K., Alexandre, F., Arildsen, T., Barba, L., Benureau, F. C. Y., Brown, C. T., de Buyl, P., Caglayan, O., Davison, A. P., Delsuc, M. A., Detorakis, G., Diem, A. K., Drix, D., Enel, P., Girard, B., Guest, O., Hall, M. G., Henriques, R. N., … Zito, T. (2017). Sustainable computational science: The ReScience initiative. *PeerJ Computer Science*, *3*, e142. https://doi.org/10.7717/peerj-cs.142

Rumsey, A. S. (2017). New-model scholarly communication: Road map for change. *Scholarly Communication Workshop*, *9*.

Saldaña, J. (2015). *The coding manual for qualitative researchers*. Sage.

Sallans, A., & Lake, S. (2014). Data management assessment and planning tools. *Research Data Management: Practical Strategies for Information Professionals*, 87–107.

Sands, A. E. (2017). *Managing Astronomy Research Data: Data Practices in the Sloan Digital Sky Survey and Large Synoptic Survey Telescope Projects* [UCLA]. https://escholarship.org/uc/item/80p1w0pm

Sands, A. E., Borgman, C. L., Traweek, S., & Wynholds, L. A. (2014). *We're working on it: Transferring the sloan digital sky survey from laboratory to library*.

Sandusky, R., Allard, S., Baird, L., Crowston, K., Forrester, A., Grant, B., Hu, R., Olendorf, R., Pollock, D., Specht, A., & Volentine, R. (2021). IJDC | Research Paper Assessment, Usability, and Sociocultural Impacts of DataONE: A Global Research Data Cyberinfrastructure Initiative. *International Journal of Digital Curation*, *16*, 1–48. https://doi.org/10.2218/ijdc.v16i1.678

Sawyer, S., Crowston, K., & Wigand, R. T. (2011). *Digital Assemblages: Evidence and Theorizing from the Computerization of the U.s. Residential Real Estate Industry 1*.

Sawyer, S., Sharma, S., Willis, M., & Østerlund, C. (n.d.). *The mess has always been with us: Organizing for distributed scientific collaboration1*.

Schmidt, K. (2002). Remarks on the complexity of cooperative work. *Revue d'intelligence Artificielle*, *16*(4–5), 443–483.

Schmidt, K. (2016). 18 Reflections on the Visibility and Invisibility of Work. *Boundary Objects and Beyond: Working with Leigh Star*, 345.

Schneider, M. V., Griffin, P. C., Tyagi, S., Flannery, M., Dayalan, S., Gladman, S., Watson-Haigh, N., Bayer, P. E., Charleston, M., Cooke, I., Cook, R., Edwards, R. J., Edwards, D., Gorse, D., McConville, M., Powell, D., Wilkins, M. R., & Lonie, A. (2019). Establishing a distributed national research infrastructure providing bioinformatics support to life science researchers in Australia. *Briefings in Bioinformatics*, *20*(2), 384–389. https://doi.org/10.1093/bib/bbx071

schraefel, m. c., & Dix, A. (2009). Within bounds and between domains: Reflecting on Making Tea within the context of design elicitation methods. *International Journal of Human-Computer Studies*, *67*(4), 313–323. https://doi.org/10.1016/j.ijhcs.2007.10.009

Scott, P., Haworth, J., Conrad, C., & Neumann, A. (1993). Notes on the classroom as field setting: Learning and teaching qualitative research in higher education. *Qualitative Research in Higher Education*, *3*(6), 3–24.

Scott, W. (2008). Institutions and Organizations: Ideas and Interests. *Institutions and Organizations: Ideas and Interests*. https://digitalcommons.usu.edu/unf_research/55

Scott, W. R. (2001). Institutions and Organizations Second Edition Sage Publications. *Inc. Thousand Oaks, Calif*.

Scott, W. R. (2008). Crafting an analytic framework I: Three pillars of institutions. *Institutions and Organizations: Ideas and Interests*.

Scott, W. R. (2013). *Institutions and organizations: Ideas, interests, and identities*. Sage publications.

Scott, W. R., Ruef, M., Mendel, P. J., & Caronna, C. A. (2000). *Institutional change and healthcare organizations: From professional dominance to managed care*. University of Chicago Press.

Scroggins, M. J., & Pasquetto, I. V. (2020). Labor Out of Place: On the Varieties and Valences of (In) visible Labor in Data-Intensive Science. *Engaging Science, Technology, and Society*, *6*, 111–132.

Sewerin, C. (2015). *Research data management faculty practices: A Canadian perspective*.

Shankar, K. (2004). Recordkeeping in the Production of Scientific Knowledge: An Ethnographic Study. *Archival Science*, *4*(3), 367–382. https://doi.org/10.1007/s10502-005-2600-1

Shankar, K., & Eschenfelder, K. (2017). Organizational and institutional work in data infrastructures. *Proceedings of the Association for Information Science and Technology*, *54*(1), 595–598. https://doi.org/10.1002/pra2.2017.14505401082

Shapin, S. (1989). The invisible technician. *American Scientist*, *77*(6), 554–563.

Shendure, J. (2008). The beginning of the end for microarrays? *Nature Methods*, *5*(7), 585–587.

Shilton, K. (2015). Anticipatory ethics for a future Internet: Analyzing values during the design of an internet infrastructure. *Science and Engineering Ethics*, *21*(1), 1–18.

Simonton, D. K. (1997). Creative productivity: A predictive and explanatory model of career trajectories and landmarks. *Psychological Review*, *104*(1), 66.

Spotlight on Qualitative Methods: Do I Need Multiple Coders? (2020, February 18). *IAPHS - Interdisciplinary Association for Population Health Science*. https://iaphs.org/demystifying-the-second-coder/

Star, S. L. (1999). The ethnography of infrastructure. *American Behavioral Scientist*, *43*(3), 377–391.

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology,translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, *19*(3), 387–420.

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, *7*(1), 111–134.

Star, S. L., & Strauss, A. (1999). Layers of silence, arenas of voice: The ecology of visible and invisible work. *Computer Supported Cooperative Work (CSCW)*, *8*(1–2), 9–30.

Steinhardt, S. B., & Jackson, S. J. (2014). Reconciling rhythms: Plans and temporal alignment in collaborative scientific work. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 134–145.

Steinhardt, S. B., & Jackson, S. J. (2015). Anticipation work: Cultivating vision in collective practice. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 443–453.

Stephan, P. E. (2012a). *How economics shapes science* (Vol. 1). Harvard University Press Cambridge, MA.

Stephan, P. E. (2012b). Research efficiency: Perverse incentives. *Nature*, *484*(7392), 29.

Stephan, P. E. (2012c). Perverse incentives. *Nature*, *484*(7392), 29–31. https://doi.org/10.1038/484029a

Stewart, B. E. (2015). In abundance: Networked participatory practices as scholarship. *The International Review of Research in Open and Distributed Learning*, *16*(3).

Stöckelová, T. (2012). Immutable Mobiles Derailed: STS, Geopolitics, and Research Assessment. *Science, Technology, & Human Values*, *37*(2), 286–311. https://doi.org/10.1177/0162243911415872

Strasser, C. A., & Hampton, S. E. (2012). The fractured lab notebook: Undergraduates and ecological data management training in the United States. *Ecosphere*, *3*(12), 1–18.

Strauss, A. (1978). A social world perspective. *Studies in Symbolic Interaction*, *1*(1), 119–128.

Strauss, A. (1982). Social worlds and legitimation processes. *Studies in Symbolic Interaction*. http://psycnet.apa.org/psycinfo/1983-20938-001

Strauss, A. (1985). Work and the division of labor. *Sociological Quarterly*, *26*(1), 1–19.

Strauss, A. (1988). The Articulation of Project Work: An Organizational Process. *The Sociological Quarterly*, *29*(2), 163–178. JSTOR.

Strauss, A., & Corbin, J. (1994). Grounded theory methodology. *Handbook of Qualitative Research*, *17*, 273–285.

Studies, D. S. (2017, September 27). *Community Level Data Science: Beyond novelty squared*. Medium. https://medium.com/@dataethnography/community-level-data-science-beyond-novelty-squared-4a70c51eae50

Suchman, L. (1995). Making work visible. *Communications of the ACM*, *38*(9), 56–64.

Suchman, L. (1996). Supporting articulation work. *Computerization and Controversy: Value Conflicts and Social Choices*, *2*, 407–423.

Suchman, L. A. (1994). Suporting Articulation Work: Aspects of a Feminist Practice of Technology Production. *Proceedings of the IFIP TC9/WG9.1 Fifth International Conference on Woman, Work and Computerization: Breaking Old Boundaries - Building New Forms*, 7–21. http://dl.acm.org/citation.cfm?id=647314.722855

Suchman, L. A. (1996). Supporting articulation work. *Computerization and Controversy: Value Conflicts and Social Choices*, *2*, 407–423.

Suddaby, R., & Greenwood, R. (2005). Rhetorical strategies of legitimacy. *Administrative Science Quarterly*, *50*(1), 35–67.

Szalay, A. S., & Blakeley, J. A. (2009). *Gray's laws: Database-centric computing in science.*

Tanjo, T., Kawai, Y., Tokunaga, K., Ogasawara, O., & Nagasaki, M. (2021). Practical guide for managing large-scale human genome data in research. *Journal of Human Genetics*, *66*(1), 39–52. https://doi.org/10.1038/s10038-020-00862-1

Tashakkori, A., Teddlie, C., & Teddlie, C. B. (1998). *Mixed methodology: Combining qualitative and quantitative approaches* (Vol. 46). Sage.

Teich, A. H. (2018). In Search of Evidence-based Science Policy: From the Endless Frontier to SciSIP. *Annals of Science and Technology Policy*, *2*(2), 75–199.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*, *6*(6). https://doi.org/10.1371/journal.pone.0021101

Tenopir, C., Allard, S., Sinha, P., Pollock, D., Newman, J., Dalton, E., Frame, M., & Baird, L. (2016). *Data management education from the perspective of science educators*.

Tenopir, C., Sandusky, R. J., Allard, S., & Birch, B. (2014). Research data management services in academic research libraries and perceptions of librarians. *Library & Information Science Research*, *36*(2), 84–90. https://doi.org/10.1016/j.lisr.2013.11.003

Thornton, P. H., Jones, C., & Kury, K. (2005). Institutional logics and institutional change in organizations: Transformation in accounting, architecture, and publishing. In *Transformation in cultural industries*. Emerald Group Publishing Limited.

Thornton, P. H., & Ocasio, W. (2008). Institutional logics. *The Sage Handbook of Organizational Institutionalism*, *840*(2008), 99–128.

Thursby, J. G., Haeussler, C., Thursby, M. C., & Jiang, L. (2018a). Prepublication disclosure of scientific results: Norms, competition, and commercial orientation. *Science Advances*, *4*(5), eaar2133. https://doi.org/10.1126/sciadv.aar2133

Thursby, J. G., Haeussler, C., Thursby, M. C., & Jiang, L. (2018b). Prepublication disclosure of scientific results: Norms, competition, and commercial orientation. *Science Advances*, *4*(5), eaar2133. https://doi.org/10.1126/sciadv.aar2133

Tolbert, P. S. (1985). Institutional environments and resource dependence: Sources of administrative structure in institutions of higher education. *Administrative Science Quarterly*, 1–13.

Tolbert, P. S., & Zucker, L. G. (1983). Institutional sources of change in the formal structure of organizations: The diffusion of civil service reform, 1880-1935. *Administrative Science Quarterly*, 22–39.

Trinkle, D. A., & Andersen, D. L. (2015). Valuing Digital Scholarship in the Tenure, Promotion, and Review Process: A Survey of Academic Historians. In *Digital Scholarship in the Tenure, Promotion and Review Process* (pp. 61–76). Routledge.

Tukey, J. W. (1977). *Exploratory Data Analysis. Pearson, Reading, Mass.* Pearson.

Tuomi, I. (1999). *Data is more than knowledge: Implications of the reversed knowledge hierarchy for knowledge management and organizational memory*. 12-pp.

Van Tuyl, S., Michalek, G., & Van Tuyl, S. (2015). Assessing research data management practices of faculty at Carnegie Mellon University. *Journal of Librarianship and Scholarly Communication*, *3*(3).

Vardigan, M., & Whiteman, C. (2007). ICPSR meets OAIS: Applying the OAIS reference model to the social science archive context. *Archival Science*, *7*(1), 73–87. https://doi.org/10.1007/s10502-006-9037-z

Venkatraman, V. (2013). When all science becomes data science. *Science*.

Vertesi, J. (2014). Seamful Spaces: Heterogeneous Infrastructures in Interaction. *Science, Technology, & Human Values*, *39*(2), 264–284. https://doi.org/10.1177/0162243913516012

Walker, A. M., DeVito, M. A., Ringland, K. E., & Reddy, M. (2019). (In)visible Choices: Articulation Work and the Rise in US Maternal Mortality. *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, 403–407. https://doi.org/10.1145/3311957.3359463

Wallis, J. C., Borgman, C. L., Mayernik, M. S., & Pepe, A. (2008). Moving Archival Practices Upstream: An Exploration of the Life Cycle of Ecological Sensing Data in Collaborative Field Research. *International Journal of Digital Curation*, *3*(1), 114–126. https://doi.org/10.2218/ijdc.v3i1.46

Walsh, L. L., Seabloom, R., & Thompson, C. W. (2017). *Range extension of the Virginia opossum (Didelphis virginiana) in North Dakota*.

Wang, P., Mathieu, R., Ke, J., & Cai, H. J. (2010). Predicting criminal recidivism with support vector machine. *2010 International Conference on Management and Service Science*, 1–9.

Weber, N. (2020). Finite and infinite games: An ethnography of institutional logics in research software sustainability. *Proceedings of the Association for Information Science and Technology*, *57*(1), e281. https://doi.org/10.1002/pra2.281

West, J. D., Jacquet, J., King, M. M., Correll, S. J., & Bergstrom, C. T. (2013). The role of gender in scholarly authorship. *PloS One*, *8*(7), e66212.

Whitlock, M. C. (2011). Data archiving in ecology and evolution: Best practices. *Trends in Ecology & Evolution*, *26*(2), 61–65. https://doi.org/10.1016/j.tree.2010.11.006

Whitmire, A. L., Boock, M., & Sutton, S. C. (2015). Variability in academic research data management practices: Implications for data services development from a faculty survey. *Program*, *49*(4).

Widmalm, S. (2016). The Practice Turn in Science Studies: Past and Present: Léna Soler, Sjoerd Zwart, Michael Lynch, and Vincent Israel-Jost (eds) (2014) Science After the Practice turn in the Philosophy, History, and Social Studies of Science. Routledge, New York and

London, ISBN: 978-0-415-72295-7, 346 pp, $145 (Hardback). *Science & Education*, *25*(7–8), 943–946. https://doi.org/10.1007/s11191-016-9854-2

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., & Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*.

Williams, J. J., & Teal, T. K. (2017). A vision for collaborative training infrastructure for bioinformatics. *Annals of the New York Academy of Sciences*, *1387*(1), 54–60. https://doi.org/10.1111/nyas.13207

Xconomy: Big Data Grant from Moore, Sloan Aims to Make Pi-Shaped Scientists. (2013, November 12). Xconomy. https://xconomy.com/seattle/2013/11/12/moore-sloan-foundations-grant-uw-berkeley-nyu-37-8m-data-science/

Yoon, A., Curty, R., Jeng, W., & Qin, J. (2016). *Untangling data sharing and reuse in social sciences*.

Zeng, M. L. (2008). *Metadata*. Neal-Schuman Publishers, Inc.

Ziewitz, M., & Lynch, M. (2018). It's Important to Go to the Laboratory: Malte Ziewitz Talks with Michael Lynch. *Engaging Science, Technology, and Society*, *4*, 366–385.

Zilber, T. B. (2002). Institutionalization as an interplay between actions, meanings, and actors: The case of a rape crisis center in Israel. *Academy of Management Journal*, *45*(1), 234–254.

CURRICULUM VITA

SARAH E. BRATT

sebratt@syr.edu ● www.linkedin.com/in/sarahbratt/
Curriculum Vitae, November 2021

EDUCATION

Syracuse University School of Information Studies Syracuse, NY
Ph.D. Student in Information Science & Technology Advisor: Dr. Jian Qin. July 2022

Syracuse University School of Information Studies Syracuse, NY
M.S. Library & Information Science, CAS Data Science May 2014

Ithaca College Ithaca, NY
B.A. Philosophy, Minors in Italian and Political Science May 2012

OBJECTIVE

Advance long-term research data sustainability by studying new ways of sharing, organizing, and collaborating made possible by information and communication technology. I approach this question in several ways: qualitatively and quantitatively for theory building and sociotechnical systems design to accelerate knowledge production and diffusion. I am especially interested in their applications to inform science policy initiatives a) promoting diversity, equality, and inclusion in science and b) addressing policy-practice gaps in research data management and c) developing improved use and impact metrics for data products and services.

Keywords: Scholarly Communication; Research Data Management; Digital Scholarship; LIS

REFEREED JOURNAL PUBLICATIONS

- Qin, J., Hemsley, J., & **Bratt, S.E.**, "The Structural Shift and Collaboration Capacity in GenBank Networks: A Longitudinal Study." *Quantitative Studies of Science* (QSS). 2022 (under review)
- Zeng, T., Wu, L.., **Bratt, S.E**., Acuna, Daniel. (2020) "Assigning credit to scientific datasets using article citation networks." *Journal of Informetrics*.
- Bandara, D., Velipasalar, S., **Bratt, S.E.**, & Hirshfield, L. (2018). Building predictive models of emotion with functional near-infrared spectroscopy. *International Journal of Human-Computer Studies*, 110, 75-85.
- Costa, Mark R., Qin, J. and **Bratt, S.E**. (2016) "Emergence of collaboration networks around large-scale data repositories: a study of the genomics community using GenBank." Scientometrics 108.1 (2016): 21-40.

REFEREED CONFERENCE PROCEEDINGS

- Hemsley, J., Qin, J., & **Bratt, S. E.** (2020). "Data to knowledge in action: A longitudinal analysis of GenBank metadata." Proceedings of the Association for Information Science and Technology, 57(1), e253.
- Hincks, S. W., **Bratt, S.E.**, Poudel, S., Phoha, V. V., Jacob, R. J., Dennett, D. C., & Hirshfield, L. M. (2020). Entropic Brain-Computer Interfaces.
- Neupane, A., Saxena, N., Hirshfield, L. M., & **Bratt, S. E.** (2019). The Crux of Voice (In) Security: A Brain Study of Speaker Legitimacy Detection. In *NDSS*.

- Qin, J., Hemsley, J., & **Bratt, S.E.** (2018). "Collaboration capacity: Measuring the impact of cyberinfrastructure-enabled collaboration networks." In: Science of Team Science (SCITS) 2018 Conference, Galveston, Texas, May 21-24, 2018.
- **Bratt, S. E.**, Hemsley, Jeff, Qin, Jian, and Costa, Mark. "Big Data, Big Metadata, and Quantitative Study of Science: A Workflow Model for Big Scientometrics." Association for Information Science and Technology (ASIST) 2017.
- **Bratt, S.E.**, Semaan, Bryan, and Franco, Zeno. "Translation in Personal Crisis: Opportunities for Wearable ICT Design." Proceedings of the 14th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2017). Albi, France.
- Hincks, S. W., **Bratt, S.E.**, Poudel, S., Phoha, V. V., Jacob, R. J., Dennett, D. C., & Hirshfield, L. M. (2017). Entropic Brain-computer Interfaces-Using fNIRS and EEG to Measure Attentional States in a Bayesian Framework. In PhyCS (pp. 23-34).
- **Bratt, S.E**. "Toward an Open Data Repository and Meta-Analysis of Cognitive Data using fNIRS Studies of Emotion." HCII 2017. International Conference on Augmented Cognition. Springer International Publishing, 2017.
- Sharma, Sarika, Sawyer, Steve, Osterlund, Carsten, **Bratt, S.E.**, and Willis, Matthew. "Theorizing Messy Work: Four perspectives on distributed scientific collaborations." GROUP 2016 Proceedings (2016).
- Costa, Mark, and **Bratt, S.E**. "Truthiness: Challenges Associated with Employing Machine Learning on Neurophysiological Sensor Data." HCII 2016. International Conference on Augmented Cognition. Springer International Publishing, 2016.
- Serwadda, A., Phoha, V. V., Poudel, S., Hirshfield, L. M., Bandara, D., **Bratt, S. E**., & Costa, M. R. (2015, September). fNIRS: A new modality for brain activity-based biometric authentication. *In 2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems* (BTAS) (pp. 1-7). IEEE.
- Hirshfield, Leanne, Costa, Mark R., Bandara, and D., **Bratt, S.E**. "Measuring situational awareness aptitude using functional near-infrared spectroscopy." International Conference on Augmented Cognition. Springer International Publishing, 2015.

MANUSCRIPTS IN PREPARATION

- **Bratt, S.E.**, Sharma, S., & Erickson, I. (2022) "*Datarticulation* Work: Enabling Genomics by Making Research Data Deposit 'Do-Able'." Association for Information Science and Technology (ASIST). (in prep)
- Qin, J., Hemsley, J., & **Bratt, S.E.** (2022). "The Role of Cyberinfrastructure-Enabled Collaboration Networks in Supporting Collaboration Capacity." PLoS Biology. *Available at SSRN 3887529*. (in prep)
- **Bratt, S.E.**, Jung, S., Semaan, B. (2022) "Designing for Non-Avoidance: Toward Legitimating New Norms for Veterans with PTSD." *Computer Supported Cooperative Work (CSCW)*. (R&R)

REFEREED POSTERS & PANELS

- **Bratt, S.E.** (2019). The Co-Production of Data Sharing Norms: From the lab to CI-enabled data repositories and back again. In *Lab Studies Reloaded? Machine Learning, Ethnography, and Critical STS. Society for the Social Studies of Science (4S) 2019 Proceedings*.
- **Bratt, S.E.**, Qin, J., & Hemsley, J. (2019). A Closer Look at Data Co-authorship: Team size trends in 'big science'. *In 17th International Conference on Scientometrics and Informetrics*, ISSI 2019 (pp. 2664-2665).
- **Bratt, S. E.,** Qin, J., Hemsley, J. J., Costa, M. R., & Wang, J. (2016). Validating science's power players: scientometric mixed methods for data verification in identifying influential scientists in a genetics collaboration community. *iConference 2016 Proceedings*.

REFEREED WORKSHOP PROCEEDINGS

- **Bratt, Sarah E.**, Qin, J., & Hemsley, J. (2021) "Analyzing data collaborations as the 'missing link' in scientific collaboration indicators using metadata analytics." Association for Information Science and Technology (ASIST)
- Qin, J., Hemsley, J., & **Bratt, S. E.** "Evaluating the Impact of Collaboration Enablers through Metadata Analytics." Proceedings of the Association for Information Science and Technology (ASIST) 2021.
- Collective, C. J., Molina León, G., Kirabo, L., Wong-Villacres, M., Karusala, N., Kumar, N., ... & Sharma, V. (2021, October). Following the Trail of Citational Justice: Critically Examining Knowledge Production in HCI. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing* (pp. 360-363).
- **Bratt, S.E.** "Unapologetically feminist scientists: Social media as resiliency-making in academic research." *Northeast Science and Technology Studies* 2020 *Graduate Student Conference* (NESTS). Rensselaer Polytechnic Institute. Troy, NY. March 5-6, 2020.
- **Bratt, S.E.** and McDonough, C. "Extending and Destabilizing the Lab." *Intersectionality in the Digital Humanities*. Syracuse University Bird Library. 2019.

FELLOWSHIPS AND AWARDS

Laboratory for Innovation Science at Harvard (LISH) Cambridge, MA
Science Production Function Society Fellow (SPFS) May 2018-August 2019

- Designed SPFS survey instruments for qualitative and quantitative data collection.
- Recruited and interviewed active R1 research scientists to refine instruments.
- Analyzed interview data using open coding methods and RQDA software.
- Conducted literature review on research groups to interpret findings.
- Wrote and edited data analysis summary reports to submit to funder.

iSchool Inclusion Institute (i3) Pittsburgh, PA
Programming Fellow ($1,000) July 2017

Syracuse University School of Information Studies Syracuse, NY
Doctoral Summer Research Grant ($6,500) June 2017

Syracuse University School of Information Studies Syracuse, NY
Recipient of the Masters' Award in Library and Information Science May 2014

RESEARCH EXPERIENCE

Graduate Research Assistant, Metadata Lab August 2014 - present
Syracuse University School of Information Studies
Project: Cyberinfrastructure (CI)-enabled Science Collaboration Network Analysis

- Analyze scientific collaboration network data from trace data of NCBI's GenBank
- Conduct exploratory and confirmatory statistical analysis, and visualization (R, Infomap)
- Apply, test, and develop team science theory to inform methodologies for scientometrics data analysis and policymaking in the science of science
- Merge data from Dataverse patent records with GenBank metadata
- Produce and distribute research findings through formal publication venues and project website
- Assist in managing a team of graduate assistants (master's level) to support project analytic goals and educational objectives

Research Practicum September 2017 – December 2017
Syracuse University School of Information Studies
Practicum advisor: Jian Qin
Project: GenBank Sequence Submission Processes and Collaborative Practices of Geneticists

- Developed an IRB amendment, instrument, consent form, and solicitation script to collect in-depth interview data with geneticists on practices related to cyber-enabled data repositories.
- Acquired and analyzed big metadata records with SQL and R to analyze macro-level properties of the interviewees' scientific collaboration networks.
- Researched automated transcription software to rapidly and accurately transcribe interview data and to analyze collaborative work practices facilitated by CI-enabled repositories.
- Published and presented results to *Social Studies of Science* (4s) panel

Research Practicum September 2016 - April 2017
Syracuse University School of Information Studies
Practicum advisor: Bryan Semaan
Project: Transition Resilience: Participatory Design of Wearables for Veterans

- Developed an information communication technology (ICT) wearable prototype using Balsamiq
- Conducted a literature review on biomedical informatics and feminist science and technology studies to inform a participatory-design study of the use and non-use of wearable technology for collaborative detection and navigation of personal transitions
- Recruited vulnerable population for a participatory design study of a smart wearable technology
- Coded transcript data and performed content analysis

Research Assistant and Neuroimaging Technician September 2014 – August 2016
Syracuse University S.I. Newhouse School of Public Communications
Media Interface Network Design (M.I.N.D.) Lab

- Designed experiments tailored to functional near-infrared spectroscopy (fNIRS) data collection.
- Conducted literature reviews in neuroimaging of emotion.
- Collected data using functional near-infrared spectroscopy (fNIRS) and EEG.
- Analyzed participant survey data and produced literature reviews of relevant empirical work.

TEACHING EXPERIENCE

Research Methods in Information Science April 2022 – June 2022 (anticipated)

- Lead lectors and lab sections for research methods in information science in a synchronous, 100% online course
- Facilitate research projects for LIS graduate students to apply to academic, public, and special library applications
- Provide feedback to students on assignments and in constructive conversation

Programming Foundations and Applications (Python) September 2019-January 2020

- Led lab and seminar in python for 25 undergraduate students
- Mentored students in office hours to foster participatory learning
- Worked collaboratively with team to develop learning outcomes and graded homework and labs

Teaching Fellow February 2017 - June 2017
iSchool Inclusion Institute (i3), University of Pittsburgh School of Computing and Information

- Co-developed syllabus material for a 2-week intensive programming module to teach underrepresented undergraduate students introduction to computer programming
- Designed and taught labs in python and R on data acquisition, cleaning, and statistical analysis
- Mentored students to support their pursuit of computing careers and/or iSchool higher education

Teaching Practicum January 2017 - May 2017
Syracuse University School of Information Studies
Course: *Natural Language Processing*

- Directed a lab for graduate students in linguistics, computer science, and information management by instructed students in POS-tagging lecture and python lab session
- Guest lectured on applied NLP in text ambiguity context of GenBank repository metadata
- Established learning outcomes connecting NLP course content to research problems

Teaching Practicum September 2016 - December 2016
Syracuse University School of Information Studies
Course*: Information Visualization*

- Designed and delivered lecture and lab for visualization and cognition
- Managed content management software (CMS) in coordination with the professor of record
- Engaged students by assisting with technical and conceptual issues with R for data visualization
- Developed student assessment (quiz and labs) to evaluate learning outcomes

GRADUATE COURSES

PhD, Information Science & Technology
Introduction to Information Science (IST 800)
Theories of Digital Technology (IST 830)
Statistical Methods in Information Science (IST777)
Social Network Analysis (IST 800)
Elicitation & Analytical Techniques (IST 800)
Theories of Information (IST 790)
Sociological Theory (SOC 611)

MS, Library and Information Science; Certificate of Advanced Study, Data Science
Data Mining
Natural Language Processing
Information Management for Information Professionals
Information Policy
Information Visualization
Information Organization & Access
Introduction to Database Administration Concepts & Management Systems
Applied Data Science
Reference & Information Literacy Services
Planning, Marketing, Assessment of Services

LEADERSHIP EXPERIENCE

Vice President, Graduate Science Policy Group (GSPG) September 2019 – December 2020
Syracuse University (SU)

- Co-create mission and vision with GSPG president and Syracuse community
- Organized social and science policy events, e.g., climate change panel, I-81 community grid debate, and graduate student policy presentations and game night
- Hosted and produced GSPG podcast https://gspg.syr.edu/podcast/

Treasurer, Graduate Science Policy Group (GSPG)
Syracuse University (SU)

- Managed GSPG budget, applied for, and administered funding for GSPG events
- Advised GSPG leadership and members on budget status to inform annual program of activities
- Researched funding opportunities to expand scope of invited speakers for climate policy series

Co-Founding Organizer, iSchool Research Day May 2018 – January 2020
Syracuse University School of Information Studies,

- Designed and organized university-wide research day to facilitate dedicated research symposium for increased inter-departmental collaboration between faculty and graduate students.
- Consult with department grant manager to organize annual instantiation of the event.

Internship Manager and Research Assistant June 2015 - August 2016
Syracuse University, S.I. Newhouse School of Public Communications

- Recruited, interviewed, and worked with a team of PIs to manage 120 interns for the Media Interface Network Design (MIND) lab over the course of 4 semesters
- Trained students on fNIRS technology, experiment design, and software (R, python).

Graduate Assistant June 2012 - May 2014
Syracuse University School of Information Studies, Employer Relations & Career Services

- Validated and analyzed graduate placement survey data and created reports for iSchool Dean of Faculty & Student Services on national graduate employment comparison
- Designed and distributed graduate employment survey using Qualtrics Survey software.
- Designed and presented competitive intelligence research and analysis to inform the development of a proposal for an undergraduate data science minor

SERVICE

PhD Student Representative September 2017 - May 2018
Syracuse University School of Information Studies, Personnel Committee

- Reviewed materials for 3rd year annual review candidates, administrative policy documents, write letters of recommendation for candidates.
- *iConference Reviewer 2016 – present*
- *ACM CHI Reviewer 2017 – 2019*
- *ACM CSCW Reviewer 2017 - present*

PROFESSIONAL ASSOCIATIONS

- Professional member the Association of Information Science & Technology (May 2016 -present)
- Professional member of the Association of Computing Machinery (May 2016 - present)

SKILLS

- R, Python, Jupyter Notebook, Map-Reduce, Hadoop, Gephi, SPSS, HTML5, MapEquation, nVivo.
- Experience with qualitative techniques (experiment design, interview, focus groups, think aloud, survey design), quantitative techniques (unsupervised/supervised ML, regression, social network analysis), and UX techniques (lab studies, participatory design, speculative design, making tea)
- Languages: English; intermediate Italian; basic Spanish; beginner Farsi (spoken)