DATA EXPERTISE AND SERVICE DEVELOPMENT IN GEOSCIENCE DATA CENTERS AND ACADEMIC LIBRARIES

BY

CHERYL A. THOMPSON

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library & Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Professor Carole L. Palmer, Chair, University of Washington
Professor Joel E. Cutcher-Gershenfeld, Brandeis University
Dr. Matthew S. Mayernik, National Center for Atmospheric Research
Professor Linda C. Smith

ABSTRACT

eScience brings the promise of advancements in scientific knowledge as well as new demands on staff who need to manage large and complex data, design user services, and enable open access. One ramification is that research institutions are extending their services and staffing to address data management concerns. As more organizations extend their operations to research data, an understanding of how to develop and support research data expertise and services is needed. How can an organization build data expertise into their staff?

This study examines how organizations develop their own data expertise and services, comparing approaches in geoscience data centers and academic libraries. Case studies of two exemplar sites are presented based on evidence from qualitative interviews and artifact collection. The case studies are extended and further informed through qualitative interviews conducted with personnel at other data centers and libraries. The study addresses how to cultivate research data expertise and staffing to support data management services. Key products include a set of expertise categories, data roles, and learning strategies. The results draw attention to the contributions that data professionals make to research projects and to ways research institutions can support data professionals and data work.

ACKNOWLEDGEMENTS

Portions of this dissertation drew from research conducted with Drs. Carole L. Palmer and Matthew S. Mayernik as part of the Data Curation Education in Research Centers (DCERC) program, funded by the Institute of Museum and Library Services (grant RE-02-10-0004-10). A second source of support for this research was my Data Share fellowship with the Research Data Alliance (RDA), funded by the National Science Foundation (grant no. 1349002). Through this fellowship, I had guidance from several mentors: Michael Witt, Chuck Humphrey, Inna Kouper, and Beth Plale.

This research would not have been possible without the wonderful guidance from my committee. Thank you to Carole Palmer for her enthusiasm, mentoring, patience, and commitment to continuous improvement plus willingness to review or chat anytime. Matthew Mayernik for his unwavering support, help making sense of NCAR, and insights that I have greatly valued throughout this project. Linda Smith for helping me process and talk through my findings, and Joel Cutcher-Gershenfeld for your continued excitement and great advice. I am grateful to all of my committee members for investing the time and their expertise into my research.

I am eternally grateful to my family for encouraging my dreams, helping me de-stress, and supporting me through the challenges. Thank you to my husband, Craig, for your unwavering support, enthusiasm about my research, and willingness to take on extra child or household work. I am grateful to my son, Cannon, for all the writing breaks plus making me laugh and listening to my thoughts on our long walks. A special thank you to my father, David, for teaching me to dream and be curious and his optimism throughout this research endeavor.

Finally, thank you to my friends and colleagues for keeping me company during my writing, reviewing drafts, providing honest feedback, and keeping me caffeinated and sane throughout this whole process.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

eScience has the potential to reveal new insights from large volumes of publicly available data and address grand challenges in our society such as economic sustainability, human welfare needs, and environmental concerns. Researchers are faced with growing expectations for open access from funders, publishers, and various stakeholders to increase returns on investment, allow for scientific verifiability, and improve the value of data. Research institutions are responding by developing data services for their scientists and user communities such as data management consulting, metadata, archiving, and data sharing. As digital data, research services, and open access expectations grow, research institutions increasingly require a workforce with new expertise, ranging from data modeling to curation, discovery, sharing, and reuse. A number of fields are emerging such as data science, data curation, geoinformatics, data journalism, bioinformatics, among others. As these new areas of practice emerge the required expertise is not well understood, and the professional boundaries are still being negotiated between information science, domain sciences, and other fields that can prepare this new type of worker.

Research institutions need an understanding of the different data roles and their expertise requirements to effectively staff these new data management services. New data-related positions and roles have emerged in the workforce such as data scientist, data librarian, data engineer, and data curator. The growth in data librarian positions has been documented in the library and information science (LIS) field (Johnston, 2017; Lyon, Wright, Corti, Edmunds, & Bennett, 2013; Maatta, 2013; Sierra, 2012); however, other sectors have reported a shortage of workers with the right skill set for working with research data (Hedstrom et al., 2015; Manyika et al., 2011; TEKSystems, 2013). eScience trends are calling for cross-cutting data staff that can understand scientific user needs, work with multiple disciplines, and advance data infrastructure

(Atkins, 2003; Lord & Macdonald, 2003; Rusbridge, 2007). These demands are similar to the foundational principles of information science (Bates, 1999) where the scientific data curators can be viewed as a new incarnation of the science librarian or scientific information professional. To fulfill workforce needs, educators need an understanding of the different data roles and their preparation requirements (Varvel, Palmer, Chao, & Sacchi, 2011). Understanding the unique expertise required for data roles is key to providing a well-trained workforce that can support eScience.

In the era of data-intensive research, organizations are extending their services to research data. The academic library literature contains many reports of new data service models (Choudhury, 2008; Dasler, Muñoz, & Nilsen, 2013; Ray, 2013; Steinhart, 2011). As libraries are exploring their role in data management on campus, many college and research libraries are proceeding cautiously in planning and offering new services for research data (Tenopir, 2014). To provide effective data services, research institutions need an understanding of the staff roles and expertise and how to build research data expertise into their organizations.

Preparing to offer research data services requires an understanding of the skill set related to data work and which skills need more emphasis. For this dissertation, I introduced the concept of research data expertise, defined as the knowledge, experiences, and practices needed to perform research data work. This expertise included the social dimensions of expertise, learned by doing the work and by participating in research data communities of practice.

1.1 Research Approach and Questions

The study examines how organizations have developed their own research data expertise, through a comparison of data centers and academic libraries. Employing a case study approach, this study examines data staffing, expertise, and service models at two research sites—one

geoscience data center and one academic library—collecting evidence from interviews and artifacts. Foundational case studies for each setting are supplemented with interviews of managers at other geoscience data centers and academic libraries to extend, enrich, and further validate the case study results.

Libraries have been growing their data services in the last decade, while research and data centers have been managing scientific data and establishing their data expertise for several decades. Given the history of data efforts, the data center community is an ideal setting to examine the process of data expertise and service development. This research investigates how data services and expertise were developed and supported in geoscience data centers and compares these findings to academic libraries. By examining organizations serving a domain community (geoscience data centers) and multidisciplinary communities (academic libraries), this study provides insights into how data expertise and services differ by context and user audience.

I investigate the following set of research questions:

1. How do organizations develop (and support) data expertise?
    a. What roles and skills emerge from this process?
2. Why do data services and staffing develop differently in each case?

The study is grounded in two case studies drawing on the following sources of data: 1) semi-structured interviews with staff responsible for data services and hiring data workers; and 2) information artifact collection of organizational reports, policies, charts, job advertisements, and related materials.

The themes of research data staffing, roles, and expertise identified in interviews and organizational documents provided insights into emerging data workforce needs and the

differences in these needs between the two research sites. The analysis of interview data contributed to the development of a set of research data expertise types and categories. I also examined the organizational structure for data services, identifying the emerging staffing arrangements and boundary spanning roles of data professionals. The interviews provided rich descriptions of how the research sites enhanced research data expertise through learning practices and conditions that promoted learning innovations.

1.2 Study Contributions

1.2.1 Intellectual Contributions

The research addresses how organizations can support research data expertise and illuminates the significance of aligning staffing and expertise for effective research data services. My study documents two models with a set of distinct elements for building data expertise into the organizational structure. The identification of a comprehensive set of data roles and expertise categories addresses how institutions align data expertise and staffing and documents the complexity of expertise depth and breadth in data professionals (Collins & Evans, 2002). Previous studies have described the skill set of data professionals as requiring breadth of knowledge areas with deep expertise in one area (Bloom, 2017; Stanton, Palmer, Blake, & Allard, 2012). This analysis revealed data professionals as needing two or three areas of deep expertise for research data work. By investigating expertise in two differing organizational contexts, the similarities and differences of the sites enhanced our understanding of research data expertise and models in configuring expertise that are crucial for organizations supporting data services and for educators preparing students for data workforce.

By investigating the development of expertise, the learning processes and conditions were revealed. The research contributes a set of learning processes that enhance expertise

development and distribution across the organization. The contributions advance our understanding of agency and institutional entrepreneurs (Garud, Jain, & Kumaraswamy, 2002; Maguire, Hardy, & Lawrence, 2004) by illustrating the different spheres of influence of data professionals for impacting organizational change and learning innovations. My study findings document the importance of local data communities of practice for advancing shared expertise and practices. This research highlights the role of data professionals in data service innovations and how they are engaging in learning to overcome research data management, sharing, and reuse challenges.

1.2.2 Practical Contributions

Approaches of how to cultivate research data expertise and services that can be adopted are important for research institution managers and administrators. The two models for building data expertise and the elements of team structure, data positions and roles, and expertise configurations provide exemplar approaches for managers planning services. Research administrators and managers can document and learn from data professionalization disconnects and dilemmas within their organization. The learning strategies can be adapted to prepare staff for working with research data and meet the demands of data-intensive research. Moreover, the identification of data roles and expertise helps educators to align data management curriculum to the workforce needs.

1.3 Structure of Dissertation

The dissertation is organized into 7 chapters. Chapter 2 situates this study in the context of the literature on eScience, data services, data workforce, and expertise. Chapter 3 summarizes the overall research design. I describe the two research sites in Chapter 4. Chapters 5 and 6 provide the analysis results on building research data expertise and learning processes. Chapter 7

presents the key findings from the analysis and concludes with a discussion of implications and future directions.

CHAPTER 2: BACKGROUND LITERATURE

Data-intensive research depends on a well-trained workforce with the right expertise to support new computational techniques and make data sets accessible and usable. The study draws on the literature from eScience, research data services, and scientific information workforce, and is informed by theories of expertise from organizational science and sociology.

2.1 eScience and Research Trends

Over the past few decades, computing and digital advances have changed how science is conducted, producing a new data-intensive phase referred to as eScience. The goals of science have always been the advancement of knowledge and discovering solutions to society's grand challenges, such as global warming, economic collapses, poverty, disease prevention, and literacy. As data volumes grow and digital technologies advance, eScience presents the potential to address these challenges through large-scale data analysis. Research is more data-intensive, collaborative, and computational, than previous eras of experimental or theoretical phases of science (Hey, Tansley, & Tolle, 2009). Thomas Kuhn described the history of science as alternating between periods of normal science and revolutionary science (Kuhn, 1996). Normal science are periods when there are normative practices for conducting research such as shared understanding of what is worth exploring, how to investigate it, and what is interesting (Kuhn, 1996). When science enters a revolutionary phase, research is characterized by revisions to scientific paradigms and practices. The current trends in eScience are shifting the paradigms and manner for conducting science, where big data and computational analyses are shattering the normative scientific practices while making it possible to study critical research areas with new methods (Hey et al., 2009; Stodden, Guo, & Ma, 2013). Historically, scientists collected data for a research question, analyzed the data, and then shelved, lost, or destroyed the data as the

researcher moved onto the next project. Nowadays, technological advances have made the long-term storage and dissemination of scientific data possible. The production and storage of digital data has been growing at accelerating rates, as more analog data are also being converted to digital formats (Kurzweil, 2004). Digital data and growth of computing networks and information systems make it possible to find, acquire, and use data from other researchers and laboratories.

Scientists are faced with growing expectations of public access to data, research products, code, and methods to increase returns on investment, allow for scientific verifiability, and increase the value of scientific data. Three White House memos required federal agencies to increase public access to federally-funded research data and products (Burwell, Vanroekel, Park, & Mancini, 2013; Holdren, 2013) and to develop access and management policies for scientific collections (Holdren, 2014). Several funding agencies and sponsors have implemented policies requiring data management plans and encouraging scientists to provide public access to data from funded research (see as examples National Institutes of Health, 2003; National Science Foundation, 2015). In addition to funders, a small but growing number of scholarly journal publishers such as Nature Publishing Group (2009) are supporting this data sharing movement by requiring or encouraging published authors to make their data publicly available. Scientists have been advocating for data sharing for over a decade in the fields of engineering (Whitbeck, 2005), earth sciences (Board on Earth Sciences and Resources, 2002), life sciences (Board on Life Sciences, 2003), economics (Anderson, Greene, McCullough, & Vinod, 2005), social sciences (Freese, 2007; Schneider, 2004), medicine (Bachrach & King, 2004), and other domains (Klump et al., 2006). A growing and diverse set of stakeholders is calling for researchers to share and make accessible their data, code, and techniques.

While sharing of research results among researchers has been part of the scientific enterprise for centuries, the new expectation of data as part of the scholarly record creates new challenges for science. In the discourse on open access, the *Journal des sçavans* in France and *Philosophical Transactions of the Royal Society* in England, both established in 1665, are often called out as significant turning points in the open exchange of scientific information. This novel approach to disseminating research results became pervasive, and there was surprisingly little change in how scientific publications functioned over the last centuries. From this perspective, the new model of open access is relatively radical. It requires scientists to share more than results, methods, and theories, but also to release data sets, code, models, and other documentation. Sharing this range of research products presents many challenges for scientists. For example, several studies have found data sharing is thwarted by lack of time, incentives, and organizational support; legal issues; concerns of misuse; technical hurdles; and lack of standards and best practices in the field (Borgman, Wallis, & Mayernik, 2012; Kuipers & van der Hoeven, 2009; Postle, Shapiro, & Biesanz, 2002; Tenopir et al., 2011; Zimmerman, 2008).

Cyberinfrastructure development is an important part of the solution to open access dilemmas. Cyberinfrastructure includes the tools, software, hardware, network, and workforce to allow dissemination, discovery, and use of digital research products. Data systems require rules, standards, policies, and staff to protect, secure, and share scientific data and products (American Council of Learned Societies, 2006; Atkins, 2003). Within the context of this study, EarthCube is a notable cyberinfrastructure project funded by the National Science Foundation (NSF) that aims to transform data management and sharing across the geosciences (Gil, Chan, Gomez, & Caron, 2014). It is just one of many cyberinfrastructure initiatives in the geosciences across the globe.

As scientists comply with new access and sharing requirements, new data tools, services

and support for these modes of research need to scale. As more research institutions offer data management and sharing services, we still have limited understanding of how organizations should build expertise and capabilities needed for eScience. The next section explores the literature on research data services.

2.2 Research Data Services

Scientific data services have been developed by research institutions to address access expectations and growing digital data collections. Research and data centers have provided valuable data products and services to their domain user communities for decades. More recently, academic libraries have been building research data management services for a multidisciplinary audience of scientists and scholars. The literature is full of examples and descriptions of new service models, software, and practices for data management, preservation, and sharing. Choudhury, Palmer, Baker, and DiLauro (2013) proposed a classification for levels of data services – data storage, archiving, preservation, and curation (see Table 2.1). The levels indicate increasing support and functionality as you move from storage to curation services. Storage is the lowest service, focused on backup and restore operations, while curation is the highest service, including discovery, reuse, and value-added services. The framework presents useful distinctions in service models for my study by distinguishing the types of assistance offered. This section reviews examples of current data services and associated challenges, focusing on the geoscience data centers and research libraries[1].

[1] Research data services and tools are also prevalent in other domains. The social sciences are well established (see as examples Altman & Crabtree, 2011; Gutmann et al., 2009), as are the life sciences (a few examples are Greenberg, White, Carrier, & Scherle, 2009; Hedges, Haft, & Knight, 2012), and humanities are also beginning to expand (such as Flanders & Hamlin, 2013).

Layers	Characteristics	Implication for PI	Implication relative to NSF
Curation	• Adding value throughout lifecycle	• Feature Extraction • New query capabilities • Crossdisciplinary	• Competitive advantage • New opportunities
Preservation	• Ensuring that data can be fully used and interpreted	• Ability to use own data in the future (e.g. 5 yrs) • Data sharing	• Satisfies NSF needs across directorates
Archiving	• Data protection including fixity, identifiers	• Provides identifiers for sharing, references, etc.	• Could satisfy most NSF requirements

Storage	• Bits on disk, tape, cloud, etc.
	• Backup and restore
		• Responsible for:
		  • Restore
		  • Sharing
		  • Staffing
			• Could be enough for now but not near-term future

Table 2.1. Levels of Data Services and Curation from Choudhury et al. (2013)

Federal and academic research and data centers in the geosciences have designed a variety of services and tools for scientific data. In particular, government agencies have been investing in access to research data and design of technologies for data solutions. For instance, the National Aeronautics and Space Administration (NASA) and National Oceanic and Atmospheric Administration *(*NOAA*)* have designed a system of distributed data centers that provide data processing, archiving, and dissemination for earth science data and other high-level products. The data centers, such as the National Snow and Ice Data Center (NSIDC) or Oak Ridge National Laboratory house satellite and field campaign data and products and provide tailored services for their user communities (NASA, 2015). In national laboratories and research centers, various data access and discovery systems collect and add value to research data for particular domain communities and types of data. For instance, NCAR, one of the research sites in this study, conducts research and disseminates atmospheric and climate sciences reference data sets, observational data, simulation code and results, and high-level data products. NCAR provides data discovery and access through several specialized large-scale systems, including the

Research Data Archive (RDA), Earth Observing Laboratory (EOL), Community Data Portal, as well as individual lab and project websites, real-time data feeds, and external repositories.

Geoscience data repositories and collections are also managed within academic institutions, such as the Center for International Earth Science Information Network (CIESIN) at Columbia University, which provides archiving, discovery, access, and mapping services for interdisciplinary research in the earth and social sciences (Downs & Chen, 2010). A study profiling 38 repositories in the atmospheric and climate sciences provides a snapshot of data services and capabilities in the U.S. Most repositories provide data access and assistance with identifiers and citation, with some degree of outreach, instruction, and user support; but less than half offer data processing, metadata enhancement, and software development services (Hou, Thompson, & Palmer, 2014). These geoscience data centers and repositories are providing valuable services and systems tailored to the domain community and its needs.

In response to faculty needs for data management and sharing, college and research libraries are designing and implementing services that serve the many disciplines represented on university campuses. Numerous reports and articles have identified the role of libraries in eScience as providing data management consulting, training, and archiving (Association of Research Libraries, 2006; Council on Library and Information Resources, 2008; Gold, 2007). More recent cases further articulate common themes in academic library data service models including data management planning, instruction, and other consulting services staffed by librarians or new cross-departmental teams for digital scholarship or data management (Bryant, Lavoie, & Malpas, 2017; Johnston, 2017; Ray, 2013). Approaches for data distribution and access in academic libraries range from institutional repositories and scientific data repositories to digital libraries (Ray, 2013). To understand national trends in academic libraries, Tenopir and

team surveyed Association of College & Research Libraries (ACRL) directors and librarians about their research data management services and plans (Tenopir et al., 2011). "Data reference" was the most common service provided by libraries, a natural extension of traditional library services (Tenopir et al., 2011). Interestingly, results from a follow-up survey indicated a decrease in libraries offering and planning to offer research data services from 2011 to 2014 (Tenopir, 2014).

Data services and expertise must be viewed in the context of challenges to managing and preserving data. A National Digital Stewardship Alliance report highlighted problems with geospatial data formats, representation, and scale in the earth sciences (Morris, 2013). Additionally, academic libraries' challenges have revolved around funding, institutional support, equipment, and staffing (Brown, 2010; Corrall, Keenan, & Afzal, 2013; Creamer, Morales, Crespo, Kafel, & Martin, 2012). Across academic library studies, staffing issues have stood out where librarians lack the right knowledge, skills, and confidence to provide data management and institutions restrict funding for data personnel (Corrall et al., 2013; Creamer et al., 2012). In addition to these staffing challenges, a study of health and science libraries found that professional territory is also being negotiated between library, information technology, and domain science units in determining where research data services should be managed (Creamer et al., 2012). This emphasis on professional boundaries further supports the importance of incorporating an Abbott lens of professionalization in this study (see Section 2.5 for a description of Abbott's theory and Abbott, 1988, 1998).

Research data services have grown and matured in data centers and research libraries with little interaction between these communities. These communities provide a unique opportunity to investigate data services and staffing from domain and multidisciplinary

perspectives. Domain and multidisciplinary perspectives also need to be considered in terms of how data expertise and services are tailored for the different types of institutions. The research communities served is an important factor in how data staffing and service models have been designed, and the knowledge and experiences that are required for data professionals to maintain and improve services within these models and build improved models for the future. The next section explores the literature on information and data workers supporting science.

2.3 Scientific Information and Data Workforce

Research support staff has been an integral part of the scientific enterprise. As early as the late 1950's, a critical need for information professionals in science was recognized. Two reports emphasized a common need for workers with a blended skill set pulling from domain and information sciences:

> There will always be a need for creators and processors of information who follow parallel paths toward the advancement of learning, but there is a new and insistent demand for a professional who understands the intellectual content of a subject, understand the principles and techniques governing its information processes, and has linguistic ability to operate effectively in both. (Cohan & Craven, 1961, p. v)
>
> … the trainee should (1) have a science or technology subject background, (2) have or get suitable experience in a library preferably of the type he intends to work in, and (3) study fundamentals common to all types of libraries preferably along with other kinds of librarians, in addition to specialties useful in his later information or documentation work. (Bonn, 1959, p. 1467)

These quotes are similar to current trends in eScience calling for cross-cutting data management

staff that can understand scientific user needs, work with multiple disciplines, and advance open data infrastructure (Atkins, 2003; Lord & Macdonald, 2003; Rusbridge, 2007).

New data positions and roles have emerged in the workforce such as data scientist, data curator, and digital librarian. The growth in data curation positions and responsibilities have been documented in the LIS field (Lyon et al., 2013; Maatta, 2013; Sierra, 2012) and other industries (Manyika et al., 2011; TEKSystems, 2013). Palmer, Thompson, Baker, & Senseney (2014) surveyed the graduates of the Specialization in Data Curation at the School of Information Sciences, University of Illinois at Urbana-Champaign, to learn about their current positions. While a majority of graduates were employed, only half were performing data curation exclusively in their positions. However, almost all of the respondents were using their data curation skills, suggesting that data curation duties are being added to established positions.

The literature is full of imprecise and inconsistent position names, roles, and definitions for data professionals (Cox & Corrall, 2013; Swan & Brown, 2008). Understanding the different types of roles and their unique expertise is key to providing a well-trained workforce that can support data-intensive science. Research centers and libraries with mature data services offer an ideal setting for exploring the new data roles and expertise as they have a history in hiring and training staff for data management.

Professional education prepares students for the new data roles and working in the field. Visionary reports have stressed the need for data curation education (American Council of Learned Societies, 2006; Atkins, 2003). Swan and Brown (2008) emphasized that "library educators have an important role to play in planning for and delivering appropriately skilled people to meet the latent demand for data librarians to manage the libraries' potential data curation role" (Swan & Brown, 2008, p. 25). However, other disciplines are interested in placing

their graduates in these new data positions. In addition to data curation programs in LIS, data education initiatives are prevalent in other fields such as computer science, domain sciences, and business. Data work are emerging professions, where claims on the right preparation and expertise have not been settled.

Professional associations and special interest groups offer opportunities for sharing data expertise. In relation to data topics, domain-specific data professional groups have emerged such as Earth Science Information Partners (ESIP) for the earth sciences, International Association for Biocuration in the life sciences, and International Association for Social Science Information Services and Technology (IASSIST) in the social sciences. Professional associations like the American Geophysical Union, American Meteorological Society, and Special Libraries Association have started special interest groups for data and informatics research. International, multidisciplinary collaborations include the Research Data Alliance, bringing together scientists, data professionals, and research institutions (Parsons & Berman, 2013).

The growth of positions, professional associations, and education provides evidence of data work as emerging professions. Cox & Corrall (2013) explored the professional jurisdiction for several academic library services including research data management. Applying a systems perspective to professions (Abbott, 1988, 1998), the study revealed how data management is an unsettled arena where academic libraries are competing with computer science and domain sciences for jurisdiction control (Cox & Corrall, 2013).

As LIS attempts to define its role in the scientific data workforce, an examination of professional jurisdiction and claims is warranted. LIS education and workforce planners need a better understanding of the types of data roles appropriate for LIS students (Varvel et al., 2011). The organization context, user community served, and professional jurisdiction and claims are

important factors in the expertise and preparation required for data professionals. The next section discusses the expertise literature as it relates to this study.

2.4 Expertise

eScience requires a workforce with expertise to organize data, understand user needs, and ensure data access and use. Expertise has been discussed for centuries, resulting in no universal definition. Going back to Greek civilization, Socrates was concerned with the experts as he discussed: "I observe that when a decision has to be taken at the state assembly about some matter of building, they send for the builders to give their advice about the buildings, and when it concerns shipbuilding they send for the shipwrights…" (Taylor, 1991, pp. 11–12). As this quote illustrates, professionals are regarded as an authority on matters in their areas of practice. Studies and theories of expertise have tried to identify what makes an expert an authority on a subject and how to develop expertise in individuals. Modern theories of expertise are grounded in two main approaches: cognitive and practice. Cognitive science has modeled expertise mostly as the accumulation of knowledge and experiences. Several studies conceptualized expertise as deliberate practice, where individuals must practice for 10,000 hours to reach expert levels (see examples of Ericsson, Krampe, & Tesch-Römer, 1993; Gladwell, 2008; Simon & Chase, 1973). The cognitive approach emphasized the accumulation of facts, theories, and experiences from an individualistic perspective, ignoring social, cultural, and historical influences in how people learn.

Grounding expertise in a practice approach integrates the social and collective nature of learning. In anthropology, the concept of community of practice portrayed learning as social participation where individuals learn by community engagement and by performing the activity (Wenger, 1998). Individuals engage in and contribute to the shared practices of a community.

These communities enculturate new members and modify their practices and knowledge over time. Orlikowski (2002) studied a global product development company, observing expertise embedded in everyday work practices. She proposed the concept of *knowing in practice* as a new approach to understanding organizational competence and expertise. This concept is explained further in the next section. The practice-based studies present a richer representation of expertise than cognitive studies as practices interweave individual and collective forces of learning.

To realize the potential of eScience, organizations are developing staff and expertise to manage large volumes of data, comply with open access expectations, and support discovery and use of research products. Several studies have identified common themes of information science skills, engineering skills, and domain knowledge required to work with data (Bermes & Fauduet, 2011; Botticelli, Fulton, Pearce-Moses, Szuter, & Watters, 2011; Kim, Addom, & Stanton, 2011; Lee, 2009; Mayernik et al., 2014; Palmer, Thompson, Tenopir, et al., 2014). Less frequently mentioned skills are working with diverse groups of people (Day, 2008; Kim et al., 2011; Lee, 2009; Mayernik et al., 2014), project coordination (Currall, Johnson, & McKinney, 2007), and being a lifelong learner (Kim et al., 2011; Thompson & Palmer, 2014) are critical to curating data in today's emerging and collaboration-intensive science. Most of the data workforce research has studied working practitioners or job advertisements, but Palmer, Thompson, Baker, et al. (2014) took a different approach by contacting data curation education graduates to understand their perspectives on data work. Data curation graduates viewed technical expertise, previous experience, and communication as important skills for data curators. Most to date studies have explored the knowledge and skills needed for data professionals using interviews, surveys, and analysis of job advertisements, but have been plagued with small sample sizes and narrow focus on one type of data work (e.g., data curation). Few studies have investigated long-

standing research institutions like geoscience data centers and were grounded in the expertise literature.

Organizations have a critical need for in-house professionals equipped with appropriate expertise to support research data management trends. My study builds on the previously identified knowledge and skill areas by examining how expertise is developed and supported in an exemplar data center and academic library. These settings have a history of building and fostering data management staff, systems, and services, making them an excellent choice to investigate data expertise and roles. Using exemplar cases, the study distinguishes which skill set is required for particular roles. Viewing expertise with a practice lens captures the social dimension missing from many current studies of the data workforce. The final section of this chapter concludes with an overview of the two conceptual frames informing this study design and analysis.

2.5 Conceptual Frame

This study draws on concepts from organizational sciences and studies of profession as the primary conceptual foundation, guiding the data collection, analysis, and interpretation. The concept, *knowing in practice*, developed by Orlikowski (2002) guides the understanding of expertise and competence needed for data work. Abbott's (1988) *professional jurisdiction* and *claims* provides an analytical lens for understanding competition in data work and professionalization issues.

Knowing in practice. Orlikowski (2002) proposed the concept of *knowing in practice* (hereafter referred to as knowing or knowing how) as a new approach to understand organizational competence and expertise. This concept focuses on the doing or active aspect of expertise: “knowing is not a static embedded capability or stable disposition of actors, but rather

an ongoing social accomplishment, constituted and reconstituted as actors engage the world of practice" (Orlikowski, 2002, p. 249). *Knowing* is embedded in work practices happening every day, over and over again. *Knowing how* is developed through practice or action. She offers the example of riding a bicycle, where a person builds their *knowing how* by actually riding the bicycle. Individuals modify their *knowing* by innovations or modifications to their practices across contexts and time. These innovations provide both individual and organizational learning, ultimately changing *knowing*. A final point that she makes is the relationship between *tacit knowledge* and *knowing* – as a person continues to engage in a practice, s/he loses the recognition of *knowing how*. Thus, tacit knowledge is one form of knowing. Her work highlights how knowing and tacit knowledge are constituted in practice and vice versa. She purposefully uses the verb *knowing* to emphasize the action or practice over the noun *knowledge* where the focus is the accumulation of facts.

Organizational expertise is developed by ongoing practices of a group of employees, which may be distributed across departments or geography. These practices in context engender a collective knowing that is continuously engaged and that enables employees to work across boundaries (e.g., professional, geographic). Orlikowski (2002) illustrates the concept of knowing in a global product development company. She employs a case study approach coupled with ethnography and acknowledges that the primary source of evidence for *knowing in practice* was gathered primarily in the interviews and documents. Given both my and Orlikowski's studies utilize the case study approach, qualitative interviews, and artifact collection, *knowing* is an ideal conceptual frame for my study. Moreover, Orlikowski's research revealed the collective nature of expertise, the power of learning by doing the activity, and how knowledge was embedded into

shared work practices. The concept of *knowing* captures the complexity of organizational knowledge, making it a useful concept to guide my understanding of expertise.

The concept of knowing has been used in studies of pharmaceutical companies (Dougherty & Dunne, 2012), telemedicine (Nicolini, 2011), and software engineering companies (Choo, 2014; Zahedi & Babar, 2014). Furthermore, Savolainen (2009) suggests knowing provides a framework for studying information use in context.

In my study, *knowing* (Orlikowski, 2002) is applied to the analysis of the expertise and competence required for data work and how *knowing* develops over time in an organization. For example, a data professional may receive a new data type for curation and embark on a series of trial and error experiments or information searches to understand how to work with this new data set. This activity requires knowing how to work with data sets and developing capabilities through learning-by-doing. Continuing education, managers supporting risk-taking, or knowledge sharing among colleagues may help the data professional in this situation. Viewing expertise as embedded in practice helps to reveal the situated and provisional nature of knowing in data management. The study collects information on work practices, information artifacts used in practices, organizational norms and conditions, and challenges in data work and strategies for resolving them. Looking at *knowing* complements the knowledge and skills already identified in the data curation literature providing a holistic picture of how to prepare students for these data roles.

Professional jurisdiction and claims. Since data work is an emerging area where many professions contribute knowledge, the concepts of professional jurisdiction and claims from Andrew Abbott (1988) enhance my understanding of the various fields contributing knowledge to data expertise. Abbott, a sociologist, examined how professions are organized to perform work

and developed his theory of *The System of Professions* (1988) using a system-view and focusing on control of work areas. Professions are "…exclusive occupational groups applying somewhat abstract knowledge to particular cases" (Abbott, 1988, p. 8). Using case studies of law, psychiatry, and librarianship, Abbott (1988) illustrated an interdependent system of professions where each profession has specific activities under their jurisdiction, the right to control a provision of services or task. Professions have various levels of control in their jurisdiction such as full control or subordination to another profession. Jurisdictional control is not permanent. External forces can abolish, create, or enhance control resulting in adjustments of jurisdiction claims in the system. Professions maintain jurisdictional control through their body of knowledge and skills needed to perform the work, ward off intruding professions, and make jurisdictional claims for new problem spaces. Abbott (1988) emphasizes the importance of looking at jurisdictional boundaries and how boundaries change as professions compete for control. Professional jurisdictions are shaped by disputes where boundaries are defined and re-defined through a series of wins and/or losses for control. Thus, a profession's story begins when a new jurisdiction emerges (e.g., data management) or another profession vacates a jurisdiction; however, competition is the key to the development of any profession.

Library and information professionals were one of Abbott's case studies in *The System of Professions* (1988). Information professionals help clients find, create, and use information. Abbott (1988) stressed librarianship as possessing only jurisdictional control over the access to cultural resources. Librarians dominate the jurisdiction of information access where they design libraries and information systems to enable the user to discover and retrieve the relevant information, in a usable format, and at the time they need it. According to Abbott (1998), librarianship is a federated profession, meaning "a loose aggregation of groups doing relatively

different kinds of work but sharing a common orientation" (p. 441). For instance, the current field of librarianship includes a variety of players such as metadata specialists, reference librarians, user experience designers, and liaison librarians all working to help users find and access information.

In LIS research, Abbott's theory has been used to understand how academic library specialties are competing for jurisdiction (Cox & Corrall, 2013); how technology has changed cataloging work (Hoffman, 2012); the development stage of information systems as a discipline (Córdoba, Pilkington, & Bernroider, 2012); how unauthorized practice of law impacts the jurisdiction of law librarianship (Trosow, 2001); and the contested terrain between the librarianship and information technology workforce (Burnett & Bonnici, 2006).

I use Abbott's concepts of professional jurisdictions and claims to investigate the work arena of data professionals and appropriate expertise for the data arena. This study captures the particular expertise contributed by data professionals and how this contribution differs from scientists, engineers, and information technology specialists. I examine the professional background, identity, affiliations, dissemination, and careers of data management staff. This study extends Abbott's case of information workers by investigating the new iteration of scientific information professionals – the data professional.

The concepts of *knowing*, professional jurisdictions, and claims guide my understanding of expertise for data work and inform my research design, providing a lens for data collection, analysis, and interpretation. I employ qualitative interviews and artifact collection to investigate these concepts in the case of data work in geoscience data center and academic library communities. The next chapter reviews the research design investigating how my sites build research data expertise.

CHAPTER 3: RESEARCH METHODS

3.1 Overview

To investigate my research questions, I developed case studies of two organizations offering mature research data services: National Center for Atmospheric Research (NCAR) and Purdue University Libraries (hereafter referred to as Purdue). The case studies were built using qualitative interviews and artifacts as evidence. Supplementary qualitative interviews with other geoscience data centers and academic libraries were used to extend the case analysis of data services and staffing at the two research sites.

The study applied a "practices approach," recognizing the "social dimension of disciplines as a primary influence on the information activities" of data professionals (Palmer & Cragin, 2008, p. 165). Situating my research to focus on data professionals allowed optimal analysis of the social unit where data practices are created and maintained (Cragin, Chao, & Palmer, 2011). The application of this "data practices" approach to the study enabled an examination of the shared knowledge, skills, and expertise of data professionals. My study's unit of analysis was the organizational unit responsible for development of research data services and curation of scientific data, acknowledging that organizations vary in the size and structure of data management teams. A practices approach emphasized the material aspects of work, prioritizing the collection of artifacts created and used during work activities. This study collected evidence from organizational artifacts, primarily documents, related to the organization's data staffing and services. Keeping with the data practices tradition of qualitative methodologies (Cragin et al., 2011; Cragin, Palmer, Carlson, & Witt, 2010), this case study integrated data from semi-structured interviews and artifact collection and employed purposive sampling.

Investigations into scientific data practices have employed a variety of methods and approaches. Surveys of scientists have reached a large number of individuals enabling comparisons across diverse groups (Tenopir et al., 2011; Tenopir, Sandusky, Allard, & Birch, 2014). This approach does not capture the details and nuances that are needed to understand practices in my study. Ethnography was another approach where researchers engage for prolonged periods in labs or teams to fully understand the behaviors of a group and producing rich descriptions of scientific practice (Borgman et al., 2012; Latour, 1999; Traweek, 1988). An ethnographic approach requires extensive access to the research sites in order to capture evidence. Given the geographic distance from my home to the research sites, a full ethnographic approach was not feasible for this study. Qualitative interviews, while used as part of ethnography, can be applied in case studies and conducted with larger samples to investigate practices, perspectives, or phenomena in a broader context, can produce rich description of my object of study, and can be conducted in a shorter timeframe, meeting the criteria needed for my study.

Through interviews, I ascertained an understanding of data expertise from the perspective of the respondents and how organizations developed their professional capacity and services. Additionally, the analysis of organizational documents revealed details about the history, mission, work arrangements, and decisions related to how organizations evolved their support for data management. The two data sources informed each other and provided multiple viewpoints.

In building the two cases based on interviews and artifacts, NCAR was the deeper case study compared to Purdue. This data center community was at the center of the study, based on my experience with preliminary work at NCAR on the Data Curation Education in Research Centers (DCERC) project. My field work at NCAR, a premier research center with multiple data

archives, influenced the premise of my research questions—that NCAR's long and productive history of managing scientific data can serve as an exemplar, offering critical insights into future needs for data expertise and data services and how they might align or differ in academic libraries.

National research and data centers like NCAR have extensive expertise in working with data and providing valuable data products to their scientific communities. The interviews with the additional set of established data centers provided a perspective on how these organizations are scaling up to manage large collections of diverse and complex data. Studying these data centers provides a platform for examining data service models, staff roles, and coordinating services that have evolved over time.

The second case of Purdue, also extended with interviews of peer institutions, allowed me to compare and explore the transferability of the data center findings to institutions where data services are in early development and serving a more diverse, multidisciplinary audience. As a case, Purdue University Libraries represented an early leader in research data service innovations in the library community. This chapter describes my study design, sources of evidence, limitations of the study design, human subjects considerations, data analysis, and data presentation.

3.2 Study Design

The study utilized a case study approach, with each case supplemented and extended with a set of qualitative interviews with peer institutions. The case study approach is valuable in that it enables an "in-depth exploration from multiple perspectives of the complexity and uniqueness of a particular project, policy, institution, programme or system in a 'real life' context" (Simons, 2009, p. 21). Though definitions of a "case" vary, a single case can be people, teams, processes,

or activities (Simons, 2009; Stake, 1995). The strengths of the approach include: deep understanding of a phenomenon; highlighting particularities of a site; documenting multiple perspectives within a site; capturing the process and dynamics of change; and flexibility of methods (Simons, 2009; Stake, 1995). Case studies have been used to investigate scholarly communication and scientific practices (such as Chen et al., 2009; Crane, 1972; Vaughan, 1999; Zimmerman, 2007), information use and user studies (see for example Fisher, Durrance, & Hinton, 2004; Veinot, 2007), and communities and their jurisdiction (examples include Abbott, 1988; Lave & Wenger, 1991).

My study built cases of two organizations with a long history of offering research data management services: NCAR and Purdue. Through case study development, I explored the complexities of data expertise development within and across these organizations. The case of NCAR was developed the most extensively due to my field experience at the site in 2014 and their long history in managing scientific data, making it an interesting case. Both cases examined the data expertise, staffing, and learning processes, allowing for comparison across organizations with domain and multidisciplinary orientations.

Geoscience data centers conduct research, manage heterogeneous and large volumes of research data, and disseminate data products and services targeted to the geosciences community. These centers include an array of organization types (federal, private, academic) and serve both the broader geosciences and related sub-disciplines including atmospheric and climate sciences. NCAR is a première, national research center in the geosciences. Academic libraries, on the other hand, represent organizations serving a multidisciplinary user community. Purdue is an ideal setting to examine as a case, given they have been designing and offering research data services to a diverse user population for over a decade.

For my study, a case was defined as the organizational units responsible for the development of research data services and curation of scientific data. These units included science teams and data teams involving a variety of workers such as scientists, data professionals, engineers, and information professionals. The research sites had multiple teams providing data services, but I chose to focus on specific teams, as described in the next section, due to the feasibility of my study. The case definition was informed by my data collection and experiences in the two organizations. Figure 3.1 presents an overview of the study design, followed by more details.

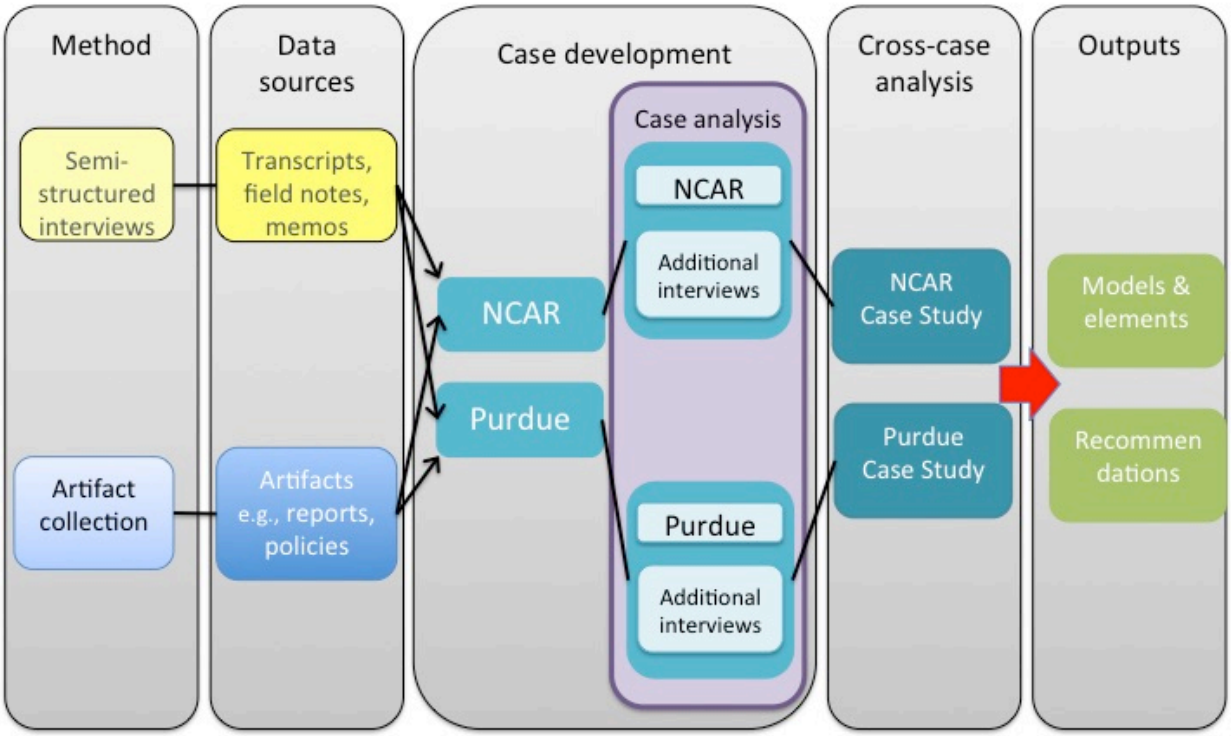

Figure 3.1. Overview of study design

3.2.1 Research Sites

A deep exploration of data expertise and service development was conducted at the two research sites. The main selection criteria for these sites were a history of offering research data

services, reputation of expertise in regard to scientific data, and ability to gain access to the site.

NCAR is a premier research center in the atmospheric sciences in Boulder, Colorado, USA. NCAR and its parent organization, the University Corporation for Atmospheric Research (UCAR), started in the 1960s and have grown today to include 7 labs and programs conducting research on such topics as sun and earth connection, air quality and chemistry, and climate systems and interactions. NCAR is a federally-funded research and development center (FFRDC), funded by the National Science Foundation (NSF), supporting research and providing instruments, field campaign support, supercomputing facilities, and archival collections of data for the atmospheric science community (Mayernik et al., 2014). While each lab has grown their data services and staff in unique ways, NCAR has a history of organization-wide efforts on data management such as the recent Data Stewardship and Engineering Team (hereafter referred to as DSET), for which I served as a graduate observer (Baker, Mayernik, Thompson, Nienhouse, Williams, & Worley, 2015). As a research center, NCAR has a long history of managing scientific data, producing valuable high-level data products, and supporting access to reference data sets, observational data, climate model code and outputs, and open access software. NCAR contains multiple data management teams across its labs. For this study, the NCAR case covers the Research Data Archive team providing data archiving and preservation support for select atmospheric and related data sets from NCAR labs; High Altitude Observatory data professionals working on large-scale observational data streams; and Climate & Global Dynamics data professionals providing data management expertise for climate modeling and simulations. These three units represent a variety of data types, atmospheric sub-disciplines, and staffing arrangements (e.g., data managers placed into research teams; data teams separated from science labs) present across NCAR.

Purdue in West Lafayette, Indiana, USA began exploring the library's role in data management services in 2004 and has evolved a range of services – data consulting, data management planning, data publishing, preservation, access, and reference. Collaborating with University of Illinois at Urbana-Champaign, Purdue designed an interview process for librarians to ascertain data curation needs from scientists as part of the IMLS-funded Data Curation Profiles project (Witt, Carlson, Brandt, & Cragin, 2009). Data curation profiles have been utilized by librarians in data management consulting at Purdue and other organizations (Brandt, 2013). This project served as a valuable learning experience for Purdue librarians to understand data practices. The experiences from the data curation profiles project were applied to the development of data management services and data education for librarians. A hallmark service is the Purdue University Research Repository (commonly referred to as PURR), offering an effective institutional approach to data publishing and sharing (Brandt, 2013). The case focused on the two data-focused teams in the library and service collaborators, liaison librarians, and library management responsible for hiring and implementing data services.

The NCAR and Purdue interview participants were data professionals and managers working in data services. Thirteen NCAR professionals participated in an interview between August and November 2015. Purdue interviews (n=10) were conducted in August to September 2015.

3.2.2 Supplementary Interviews

To extend and understand the broader applicability of the case study results, interviews were conducted with data services personnel at geoscience research/data centers and academic libraries. These sites served to validate the research site findings, determine the transferability of the results to the broader community, and to understand which aspects of the case may be unique

to these sites. For the data centers, sites were selected that represent a range of organization types, data service models, and communities served. However, the academic libraries selection represented organizational structure for data services, public/private status, and USA geographic regions.

To complement NCAR, I used a set of interviews with geoscience research and data center managers across the country that I conducted as part of the DCERC project. The evidence from these interviews provided views on the expertise needed for workforce development, services provided, and sustaining infrastructure and services in large-scale data centers (Thompson, Mayernik, Palmer, Allard, & Tenopir, 2015).

The sample of data centers was identified using the list of federally-funded research and development centers (FFRDCs), the attendee list from the EarthCube Data Facilities meeting (January 15-17, 2014), and prominent centers known to the DCERC team. Of the 32 centers identified, 21 were selected to represent a variety of geoscience domains, organization types (e.g., government, academic, non-profit), and primary activities (e.g., data services, research and development). Twenty interviews were completed between June and October 2014. Interviews averaged an hour in duration. For this study, I selected 18 of these interviews to validate the results from NCAR because these 18 centers performed research and data access as their primary activities.

To extend the Purdue case, a set of supplementary interviews were conducted to explore data services and expertise in peer academic libraries as part of my Research Data Alliance Data Share Fellowship. Complementary to the DCERC interviews, managers and staff working on research data services in libraries were interviewed to collect information on the library operations and staffing for scientific data management, data services offered, required

knowledge and skills, and history of data initiatives. The sample of academic libraries represented different organizational approaches to data services, private/public status, and countries. An initial sample of 23 libraries was selected to recruit for my fellowship study. Twenty-two interviews were completed between November 2015 and February 2016. For this research, I analyzed 14 of these interviews to validate the results from Purdue because these libraries were located in the USA.

3.3 Data Sources

For my study, the two data sources were qualitative interviews and artifacts. This multiple method approach allowed for cross-checking and cross-validation of the findings using different data sources.

3.3.1 Semi-structured Interviews

Evidence was collected from semi-structured interviews. These interviews captured information on research data services and operations, organizational structure, roles and responsibilities, necessary knowledge and skills, and professional backgrounds and identities. Since the set of DCERC interviews were conducted prior to the start of my dissertation research, I modified slightly this schedule of questions for the NCAR data collection. The language in the NCAR schedule of questions was modified for use with the academic library sites. See Appendix A for the interview questions. NCAR and Purdue interviews targeted data professionals, managers, scientists, engineers, and librarians involved in the data services of the selected teams, while supplementary interviews recruited only one or two informants at each site that could address the research questions.

The semi-structured interview technique provided me with planned questions to guide the conversation while allowing flexibility in exploring related and interesting topics related to the

research problem. The technique was ideal for my study as it allowed me to gain an in-depth understanding of the participant's perspective on data expertise and services but also captured unplanned, relevant topics.

Participant Recruitment and Selection: In qualitative research, sampling is not intended to support inference from representative samples to a larger population (Onwebguzie & Leech, 2005). Qualitative methodology and case study approach aim for sampling of cases that are information rich, addressing the study's research questions and maximizing learning (Patton, 1999; Stake, 1995). The purposive sampling technique produces quality data that is detailed and relevant to the research topic (Creswell, 2009). My study involved purposive sampling to select participants that could offer a perspective on the research questions on data services, staffing, and expertise and who were more knowledgeable informants able to address my interview questions. Participant selection focused on professionals who are responsible for working with scientific data, planning data services, or hiring staff for data services. This sampling technique included multiple groups within the research site to assess the reliability of the account on data expertise and learning processes.

Across the research and supplementary sites, there were a total of 55 participants in the study. See Table 3.1 for a summary of participants by site. I conducted 59 interviews involving participants with diverse position titles such as Project Scientist, Associate Professor, Software Engineer, Associate Dean, among others.

Site	Number of participants	Number of interview sessions	Example position titles
NCAR	13	15	Project Scientist, Software Engineer
Purdue	10	12	Associate Professor, Dean of Libraries, Software Engineer, Data Curator
Supplementary data centers	18	18	Scientist, Database Engineer, Manager
Supplementary libraries	14	14	Data Management Specialist, Data Librarian, & Research Data Services Director

Table 3.1. Summary of participant sample for research and supplementary sites

Interview Process: In interviews, the researcher needs a prepared mind in order to ascertain quality evidence. For each interview, I ensured that I had an understanding of the organization and unit, mission, data services, topics in the discipline's literature, and current trends in the data curation literature. This prepared mind approach allowed me to optimize the interview time with participants.

Framing strategies were applied prior to the interview session. This strategy is important to establish study context and reduce ambiguities (Briggs, 1986). Beginning with recruitment and through all study communication, I informed potential participants of the study questions and scope and explained the topics of interest and goals of the data collection. The study communication and language was tailored to fit the setting (e.g., data center, academic library). My interview questions included dynamic and thematic question types designed to build rapport and trust while at the same time obtaining information relevant to my research (Kvale, 1996).

Qualitative studies aim to represent meaning, experience, and local knowledge. Understanding the interviewee's meaning is a process of "…looking, asking questions, and paying attention to what is relevant to people in some indigenous groups. But the key process lies in sensitively representing in written texts what local people consider meaningful and important" (Emerson, Fretz, & Shaw, 2011, p. 129). It was essential to capture local

organizational practices and meanings to understand expertise and services. Fieldnotes were employed to record interview details and reflect on my perceptions, ideas, and hunches. After the interviews and site visits, I wrote fieldnotes and memos as a means to understand local meanings and practices and reflect on what I was learning.

Prior to the interview, informed consent was obtained from each participant and I ensured that participants understood their rights as research participants. Interviews were conducted in a variety of modes to accommodate the participant's needs. If possible, participants were interviewed in person and in their workspaces. Additional modes of interviewing included conference calls and video calls for convenience. All interviews were digitally recorded and transcribed.

3.3.2 Artifact Collection

To build the cases and verify the interviewee's account, information artifacts were collected as evidence. These artifacts included organizational mission statement, policies, job advertisements, organization charts, technical reports, webpages, publications, and other material that illustrate data services, expertise requirements, or data practices. I started by searching the organization website and institutional repositories for artifacts and then solicited additional documents from interview participants.

From NCAR, I collected a set of artifacts about the three data services units in this study and organizational level data efforts. These artifacts included job advertisements, position descriptions, staff newsletters, oral histories, reports, and webpages. The organization webpages contained the history of NCAR, organizational structure, policies, mission statements, data service descriptions, and reports documenting research projects and data initiatives. From the NCAR institutional archive, I harvested staff newsletters, technical reports, and job postings.

Staff newsletters, historical summaries, and reports contain information on changing organizational priorities, services, and the restructuring of labs or units. From my analysis, I observed the job advertisements and position descriptions were not reflective of all the actual duties and responsibilities of data professionals. There were differences in the amount of documentation that was available for the three teams and for different time periods of NCAR's history.

At Purdue, I obtained artifacts from participants, website, and scholarly literature. From the website, I harvested the organization chart, library policies and mission statement, library history, data service descriptions, job advertisements, library statistics on staff and services, data software information, and library staff background. From interview participants, I collected the data service budget, position descriptions, and reports. A final artifact source was the Library and Information Science literature where Purdue staff published conference papers, journal articles, and book chapters documenting their data services and lessons learned. These documents were used to verify interview participants' accounts on data roles and expertise and provide an organizational perspective on research data services and priorities.

3.4 Limitations of Study Design

Limitations of this study design were generalizability, limited data on academic libraries, small sample, and my limited knowledge of the geosciences. Despite generalizability being a common critique of the case study approach and qualitative methods, I selected my methods because: 1) my research questions were aimed at the site level; and 2) I wanted to capture the complexities and particularities of the sites. The set of supplementary interviews, combined with evidence from the literature, allowed me to understand how unique my sites were in terms of peer institutions and how transferable my results were to similar organization types.

A common caveat in multiple case studies is one case will be more developed than the other cases. The deeper case of NCAR in comparison to Purdue was a second limitation. The study was designed to explore data services and staffing in academic libraries, distinguishing service models and evaluating the applicability of data center approaches. The supplemental libraries were selected to represent different types of organizations and staffing models, allowing me to optimize my learning. My case results offered a foundational exploration of service models and staffing expertise in libraries for future studies to compare their results.

The case studies were limited to only two research sites. NCAR and Purdue were selected for this study due to their history with data management services and their domain and multi-disciplinary perspectives. The two sites do not encompass the diversity of data services and staffing approaches in research institutions. The sample of data professionals and teams may not provide a comprehensive assessment of research data management services in these organizations. At NCAR, the teams were selected to represent different data types and sub-disciplines of atmospheric science but do not represent the diversity of research across NCAR labs. At Purdue, I focused my study on the data services provided by the library and did not capture services provided in disciplinary departments or research centers. The study of long-standing research institutions and their data services is an area for continued exploration.

A final limitation was my limited knowledge of the geosciences. Through my NCAR field experience and research of geoscience data centers, I have gained knowledge of atmospheric and climate sciences and their data issues. During this study, I relied on my background in information science and social science data management for an understanding of data management and curation concepts. I also used the geoscience literature and committee

members with atmospheric science expertise to gain an understanding of concepts, research techniques, and data practices.

3.5 Human Subjects Review

The research approach for working with human subjects was approved by the University of Illinois at Urbana-Champaign Institutional Review Board (UIUC IRB). The involvement of NCAR staff members required additional review and approval by the NCAR Review Board. The study complied with technology and human subjects regulations and practices for safeguarding the data. See Appendix B for study approval letters from UIUC and NCAR review boards.

3.6 Analysis

I employed qualitative analysis to address my research questions. Transcripts, fieldnotes, and artifacts were imported into ATLAS.ti, qualitative analysis software. ATLAS.ti allowed the addition of document attributes, where I indicated the case, site, and interviewee demographics. Analysis was conducted in two phases – 1) case analysis and 2) cross-case analysis – explained in more detail below. Freewriting memos and mapping were conducted as initial analysis techniques to unlock memories, make connections, and organize ideas (Sustein & Chiseri-Strater, 2012). Codes were developed both inductively and deductively. Open coding identified all the themes, ideas, and concepts in the transcripts and documents (Emerson et al., 2011). Then, I drafted a set of codes from the concepts of knowing in practice, professional jurisdiction and claims, levels of data services, data practices, and relevant topics in the data curation literature. A set of final codes was selected for focused coding (Emerson et al., 2011). The final codes covered themes of data expertise services, staffing, and profession, as well as organization and research characteristics. See the codebook in Appendix C.

3.6.1 Case Analysis

Case interviews and artifacts were used to develop the cases of NCAR and Purdue, using the supplementary interviews for comparative analysis. The case analysis focused on the identification of data expertise and service models, specifically the roles, knowledge, and systems needed to offer data services at each site. A secondary aim was to understand the organizational history and characteristics that led to the development of data staffing and services. The analysis considered the knowing in practice and learning processes articulated by staff in interviews and appearing in organizational documents. A final consideration of this analysis was the professional backgrounds and identities of the participants, informing professional jurisdiction and claims for data work. This analysis addressed my first research question on how each research site built research data expertise and services.

Individual case reports for Purdue and NCAR were developed from the qualitative analysis and coding. The initial codes were tested on a subset of interviews and documents from the two research sites. From this coding test, I refined code definitions and identified new analytical themes. A final set of codes was produced and applied to all interviews and artifacts, including re-coding the test set. ATLAS.ti was employed for coding and tracking the coding list and definitions. The case reports were constructed by gathering all the coded transcripts and artifacts. Evidence for the following concepts was pulled from the code reports:

- Education background, knowledge, skills, and experiences necessary or preferred for scientific data management staff;
- Expertise;
- Personal and organizational learning related to data management;
- Roles, responsibilities, and organizational structures for scientific data management;

- Services, data products, software, and systems for scientific data management and tailoring to user communities;
- Professional backgrounds, identities, and association memberships of data workers;
- Professional jurisdiction, disputes, and claims;
- Organizational development processes, enablers, barriers, and pivotal events in offering data services; and
- Other emergent themes from coding (e.g., invisible work).

Mapping and memos documented my learning – findings, observations, and interpretations for each case. Through comparative analysis of the research site and set of supplementary interviews, I compared the site with peer institutions based on the examination of data services, staffing, expertise, and organizational characteristics.

3.6.2 Cross-case Analysis

The cross-case analysis focused on the commonalities and differences between NCAR and Purdue. NCAR served as the anchor case to compare the findings from Purdue. Analysis across cases examined differences in the data expertise and learning approaches. The second aim of this analysis was to explore explanations for why certain services models, expertise, roles, and staffing structures developed in each case.

This analysis addressed my second research question on why services and staffing developed differently in these organizational contexts. In Chapter 5, the results discussed what aspects of data expertise and staffing are similar and different across cases. This analysis produced a set of data roles and expertise categories observed in both research sites.

Data expertise types development: The interview data and documents contributed to the development of data expertise types. The interviews were the primary source for this analysis,

providing a thick, rich description of the learning, staff background, and data practices embedding knowledge. Types of expertise were identified by first exploring the NCAR data and, second, comparing to the Purdue data. The NCAR data produced 12 types of data expertise. Using the Purdue data, these 12 types were validated and a set of 6 additional types was identified. I revisited the NCAR data to verify the presence of these 6 types of expertise. Eighteen types of research data expertise are identified and described in Chapter 5.

3.6.3 Strategies for Validating Results

The case results were validated in two phases. The first phase utilized the integration of data sources and thick description as strategies for validating the case results of NCAR and Purdue. The integration of data sources provided multiple perspectives and a comprehensive account of data expertise development (Lincoln & Guba, 1985; Patton, 1999). Thick description enables the readers to make transferability judgments about my findings (Creswell, 2009; Lincoln & Guba, 1985). In Chapter 4, I provide detailed description of the settings to help contextualize the results in Chapters 5 and 6.

The second phase of case development included comparative analysis of the research sites to the supplemental interviews. I examined the similarities and differences between my research sites and the peer institutions represented in the set of supplemental interviews. The research sites and supplemental sites represent a diversity of data services, staffing, partnerships, and funding sources. See Tables 3.2-3.3 for a profiling of the characteristics of supplemental data centers and libraries. With the NCAR case, there were strong similarities with 6 long-standing, federally-funded research centers and to 4 research centers with an innovative spirit in the sample. Purdue, as a site, stood out from the sample of supplemental libraries. Only one library exhibited a similar entrepreneurial spirit and sophisticated data staffing and service model to

Purdue. I observed a variety of data staffing approaches in the supplemental libraries, and the Purdue service model was confirmed as one trend in library data staffing. The supplemental libraries confirmed the trends in boundary spanning data positions, multiple expertise areas and levels, and some of the learning strategies. I have reported additional results in Chapter 5 for the supplemental libraries for areas of data staffing models and expertise where differences existed.

Characteristics	Categories	Number of Supplemental Data Centers (n=18)
Primary activity	Data services only Research only Research & development Research & data access Research & education	5 1 2 7 3
US geographic region	Northeast Midwest South West	4 1 7 6
Organization type	Federal government Non-profit Research consortium State government University or university consortium	5 3 4 1 5
Primary funding source	City government State government Department of Energy Department of Interior NASA NOAA NSF	2 1 2 1 3 3 6
Primary domain	Earth science Ecology Environmental science Geology Geophysical Glaciology Oceanography Space technology Urban science Weather	1 5 1 2 1 1 3 1 2 1
Data staffing model	Data team only Science team only Data & science teams	9 5 4
Typical data positions	Engineers, Scientists, Data managers, Archivist	
Typical data services	Data management, storage, preservation, metadata & identifiers, derived data products, dissemination, discovery	
Typical partnerships	External data producer or repository	

Table 3.2. Characteristics of supplementary data centers

Characteristics	Categories	Number of Supplemental Libraries (n=14)
US geographic region	Northeast Midwest South West	2 5 5 2
Big Ten library	Yes No	5 9
Public/Private status	Public Private	9 5
Primary funding source *(not mutually exclusive)*	Library University level Research grants	13 4 2
Data staffing model	Solo librarian without support Solo librarian with support Data team Library team with other functions Nascent/Still emerging	1 3 4 4 2
Typical data positions	Data librarian, Liaison librarian, Repository developer, Digital scholarship librarian, Metadata specialist	
Typical data services	Instruction, data consultations, data discovery, metadata, preservation, licensing and intellectual property rights, libguide	
Typical partnerships	campus information technology department, sponsored research office	

Table 3.3. Summary of supplementary libraries characteristics

In building my two cases, the supplementary interviews enabled me to compare several of the NCAR and Purdue findings to results at the additional data centers and libraries and determine the transferability of the case results to the broader community. Through the investigation of similarities and differences, I was able to gain more confidence in my findings on research data expertise and data position trends and understand areas where more variation may exist in the broader community, such as with staffing approaches to data services in academic libraries. I was not able to assess the transferability of my Chapter 6 findings given the set of supplemental interviews was collected for other studies and did not cover the topics of

organizational learning and expertise development. For future research, the application of the results from my sites to other research institutions should be done with caution, treating my findings as indicative of trends and concepts. Methodological description in Chapter 3 and contextual information in Chapter 4-6 have been provided to help the reader determine how findings might transfer to other organizations.

3.7 Data Presentation

In this dissertation, names of interviewees have been replaced with a general position title describing their role in order to protect confidentiality of the participants while making it possible to discuss the findings. The gender identification, academic majors, and career histories have been concealed to provide as much anonymity as possible for participants. Historical figures, not interviewed, are named if his/her name appeared in archival material.

In reporting the results, I use quotes from participants to illustrate the concepts and themes. To protect the confidentiality of the participants, I assigned anonymous identification numbers. To avoid confusion about which site the participant belongs to, I use two abbreviations with the quotes – first part contains the site name (i.e., NCAR or Purdue) and second part contains the participant's identification number. For instance, NCAR 201 was an NCAR participant and Purdue 101 was a Purdue participant.

I employ several conventions for presenting participant's quotes in the dissertation:

- Square brackets [ ] are used to add information that is not present in the original words of the participant to improve the readability of the excerpt. For instance, a participant may use the word, *it*, to refer to the organization and I modify the quote to include the additional term (i.e., "it [NCAR]").
- Ellipses … means I have taken out words or sentences from the original excerpt to improve clarity or remove irrelevant information.
- [Participant name redacted] is used to indicate that I have removed the participant's name to protect confidentiality.

CHAPTER 4: RESEARCH SITE PROFILES

The research developed case studies of National Center for Atmospheric Research (NCAR) and Purdue University Libraries (Purdue). These organizations have been early innovators in research data services, providing a unique opportunity to observe how data expertise has developed in two different contexts, a national research center and academic library. This chapter profiles the two sites, illustrating details and events related to research data management and providing context for the following chapters of results.

Although the data collection occurred in 2015, this research captured historical and current snapshots of each organization. The case profile uses past tense to reflect historical snapshots and present tense to indicate the current snapshot in 2015. Both organizations are still in operation at the time of this writing; however, organizational priorities, staff, services, and other aspects may have changed since the data collection.

Each profile is divided into 6 sections: organizational overview, staffing, data efforts, data services, external drivers, and timeline of key data related events. These sections correspond to elements of the research questions and analysis themes.

4.1 NCAR Profile

The National Center for Atmospheric Research (NCAR) is a national research and data center in the atmospheric and climate sciences. Since its founding in 1960, scientists and research support staff have been collecting and managing scientific data and products. As previously mentioned, this study looked at three teams at NCAR. This profile focuses specifically on the three teams and provides some organization-wide context for the study results.

4.1.1 Organizational Overview

As a respected research center in the atmospheric sciences, NCAR conducts innovative research, education, and services with the mission "to understand the behavior of the atmosphere and related Earth and geospace systems..." (University Corporation for Atmospheric Research, 2017a). Founded in 1960 in Boulder, CO, USA, NCAR is a federally-funded research and development center (FFRDC) with sponsorship from the National Science Foundation and managed by the University Corporation for Atmospheric Research (UCAR), a non-profit consortium of universities focused on atmospheric research and education. See Table 4.1 for a summary of organizational characteristics such as activities, domain, staffing, and structure. As a national center, NCAR and its labs strive to stay ahead of their peers by producing cutting-edge research, data products, services, and technologies for the geoscience community and general public.

Short name	NCAR
Org Type	Federally funded research and development center (FFRDC)
Primary activities	Research and development, data access
Domain	Atmospheric and climate sciences
Primary funding	National Science Foundation
Location	Boulder, CO
Founding year	1960
Staff size	About 1,300 staff in NCAR/UCAR in 2016
Labs	• Atmospheric Chemistry Observations & Modeling (ACOM) • Climate & Global Dynamics (NCAR-Climate) • Computational & Information Systems Laboratory (CISL) • Earth Observing Laboratory (EOL) • High Altitude Observatory (NCAR-Solar) • Mesoscale & Microscale Meteorology Laboratory (MMM) • Research Applications Laboratory (RAL)

Table 4.1. Organizational characteristics of National Center for Atmospheric Research (NCAR)

At the time of this writing, NCAR is organized into seven labs classified in two manners by sub-disciplines of atmospheric science (e.g., climatology, atmospheric chemistry) and by function (e.g., computing). NCAR as a whole is a partially centralized organization allowing individual labs the flexibility to configure teams, roles, and work to meet their scientific goals. In 2011, the NCAR Library hired the first and only Research Data Scientist to provide cross-unit services related to data management and curation.

NCAR labs contain multiple configurations for data management staffing from data-focused teams of data professionals to science teams including a data professional. My study focused on three, diverse teams that represent the variety of staff configurations and research activities:

- Research Data Archive (NCAR-Archive) is a data archiving team in Computational & Information Systems Laboratory (CISL) serving select atmospheric and climate data sets generated by NCAR and external agencies,
- Climate and Global Dynamics lab (NCAR-Climate) places data professionals into

modeling and simulation teams, and

- High Altitude Observatory (NCAR-Solar) locates data professionals into observing instrument teams.

The three teams present a variety of services, data types, and staffing observed at NCAR.

4.1.2 Data Staffing Approaches

NCAR and its labs have evolved a variety of approaches to placing data expertise next to the science. Of the teams I studied, data professionals were present in data-focused teams and science teams. Some science teams at NCAR do not have a data management expert or a staff member who identifies as a data professional. This profile reviews the three teams that I observed: NCAR-Archive, NCAR-Climate, and NCAR-Solar. Table 4.2 is a comparison of the three teams including primary activities, data staffing, services, and holdings.

Short name	NCAR-Archive	NCAR-Climate	NCAR-Solar
Full name	Research Data Archive, Computational & Information Systems Laboratory	Climate and Global Dynamics Laboratory	High Altitude Observatory
Primary activities	Data archiving; data access	Modeling and simulation; data management & access	Observational research; data management & access
Domain	Atmospheric and climate sciences	Climate sciences	Sun-earth interactions
Funding	NCAR	NCAR, Department of Energy	NCAR
Founding year	1965	1987	1940, merged into NCAR in 1960
Primary data personnel	Software engineers	Software engineers, project scientists	Software engineers, project scientists
Primary data service activities	• Data archiving • Metadata generation • High-level data products generation • Archival collection development • Data dissemination	• Data/model runs • Metadata generation • Data/model dissemination	• Data processing • Metadata generation • High-level data products generation • Data dissemination • Data visualization
Holdings scope	• NCAR/UCAR data and related information • Reanalysis data sets	• Lab model code and outputs	• Lab data products

Table 4.2. Comparison of three NCAR teams

NCAR-Archive is a data-focused team of data professionals working on data archiving and sharing for all NCAR labs. At the time of data collection, the team includes 8 Data Engineers (official title is software engineer) led by a Data Service Manager and a Senior Data Engineer with additional supervisory responsibilities.

The team has evolved its staff roles and strategies over the years. In 1965, the unit began as the Data Support Section (DSS) under the leadership of Roy Jenne with the mission to support NCAR scientists in data archiving, sharing, and programming (Jenne, 2005). In the early years, the group served all NCAR labs utilizing a "one data specialist to one scientist" strategy (NCAR

211). Early observational and climate model data required manual processing, storage on punch cards or magnetic tapes, and data sharing via printed pages and tapes distributed by postal mail or in-person. This team started with two staff members, and data work was limited. Jenne (2005) noted that, *"*with hundreds of millions of observations, it is clear that the amount of manual intervention involved in the cleanup process must be limited…to the point that the data can be easily used" (Jenne, 2005, p. 2). An early hallmark achievement of this team was the production of NCEP/NCAR global atmospheric reanalysis data sets for the earth science community.

With the retirement of Jenne, a data professional in the team assumed leadership in 2003 becoming the Data Service Manager and renamed the unit to the Research Data Archive. Since the new manager was an internal candidate, s/he had an intimate knowledge of the data work, strategies, and challenges of this team. Coupled with the change in leadership, computation advances enabled new methods for data dissemination (e.g., FTP and web downloads) and automation of routine processes. To take advantage of computational efficiencies, the new manager incorporated more technologies in the unit's work (e.g., FTP; relational databases, code libraries) moving the service strategy from a "one data specialist-to-one scientist "to a "one specialist-to-many scientists" (NCAR 211) where data professionals were responsible for specific data types (e.g., lidar) and/or sub-disciplines (e.g., oceanography). The new management also re-designed the data positions to allow for professional development. The Data Engineer position included the duties of data curation, engineering, and a special project that represented an opportunity for individual learning and service enhancements. NCAR-Archive represents a data archiving team with a long history of providing data services and evolving to meet the demands of the atmospheric research community.

NCAR-Climate laboratory is a science-focused lab with the mission to "discover the key

processes in each component of the Earth's climate system and the interactions among them" through cutting-edge research, modeling, and data services (University Corporation for Atmospheric Research, 2017b). To accomplish this mission, the lab partners with other NCAR units, government agencies, and universities. This lab traces its origins back to the NCAR's Atmospheric Analysis and Prediction Division, where in 1987 the division shifted the focus to global- and climate-level predictions, assumed a new name of Climate and Global Dynamics, and changed leadership to Warren Washington, a distinguished and nationally-recognized climate scientist.

In 2007, the current Lab Director assumed leadership, rising from the ranks where s/he was a climate scientist. The lab is divided into teams organized primarily by model types. Over the history of this lab, data professionals have been placed into modeling and simulation teams. The strategy of one data professional to one project has been a common theme in this lab. Often, the data professional assumes a Data Scientist or Data Engineer position. All data professionals were supervised by scientists or engineers that do not understand data management and services. This division has a few seasoned and experienced data professionals that have worked there for decades. The staffing approach at NCAR-Climate differs from the NCAR-Archive approach, offering an interesting opportunity to see how the placement of data professionals into science teams impacts the research data expertise requirements.

NCAR-Solar investigates the sun and earth connections providing innovative research, education, advocacy, and data services. The lab produces observational data from large instruments such as the Mauna Loa Solar Observatory (MLSO) and simulations predicting solar and upper atmosphere interactions. The lab has a longer history than NCAR. It was founded in 1940 as a small observatory in Climax, CO, USA started by Walter Orr Roberts and Donald

Menzel of Harvard College Observatory. The observatory was merged into NCAR in 1960 as an agreement for Walter Orr Roberts to become the first director of NCAR. In 1965, the lab established the Mauna Loa Solar Observatory, moving their instruments from Climax to Mauna Loa, Hawaii. Three active instruments are located at Mauna Loa, collecting data routinely (e.g., daily, monthly) and for special events or projects. The current NCAR Deputy Director was the director of NCAR-Solar until 2014, when an internal scientist stepped up as Lab Director.

At the time of data collection, NCAR-Solar staff is organized primarily into instrument teams. While most teams work with observational data streams from one instrument, one team focuses on extending climate models and simulations to the upper atmosphere. Teams are comprised of scientists and research support staff (e.g., instrument operators, software engineers). Similar to NCAR-Climate lab, data professionals historically have been located in science teams as Data Engineers or Data Scientists, holding traditional positions of software engineers or project scientists with the additional responsibilities of data management. The data professional serves one instrument or project team, focusing their curation efforts on one data stream. For the three instruments at MLSO, two Data Scientists and one Data Engineer are responsible for scientific data management for these data streams plus another Data Engineer assists with web design and visualizations for the online data catalog. To continue to provide innovative data products and services, data professionals have needed to continually update their skills, and a senior data professional working on MLSO has implemented staff training for new data professionals.

4.1.3 NCAR-wide Data Efforts

While each lab has their own data initiatives, NCAR has a history of organization-wide efforts on data management (Baker, Mayernik, Thompson, Nienhouse, Williams, & Worley,

2015). Previous data initiatives included the Information Infrastructure Technologies Applications (commonly referred to as IITA) starting in 1995, the Data Management Working Group (commonly referred to as DMWG) in 2001, and the Data Citation Working Group (hereafter referred to as DCite) in 2011. The IITA was a formally-recognized and funded group with the mission to improve discovery, access, and use of UCAR data, software, and other products. I was unable to learn much about this group and their outcomes due to limited documentation. DMWG was a sub-group of the UCAR Information Technology Council, formed to improve interoperability of scientific computing systems. A major outcome of this initiative was the NCAR Community Data Portal. This online catalog was available but not maintained during my data collection; there were plans to migrate the holdings to a new system. DCite was formed by the NCAR library staff and data archive managers growing to include a larger group of NCAR/UCAR staff. The mission of this group was to create an organization-wide approach to Digital Object Identifiers (DOIs) assignment. A key product of this group was a technical report with a set of recommendations for data citation and identifier practices, archived in the NCAR institutional repository (Mayernik et al., 2012).

A recent and continuing effort is the Data Stewardship and Engineering Team (DSET) aiming to develop, promote, and adopt organizational best practices for data and an NCAR-wide discovery system. DSET, started in 2014, is comprised of data representatives from all NCAR labs. This initiative has support from the NCAR director and dedicated resources (e.g., staff time). Early key products include an inventory of NCAR digital assets and documentation of metadata practices assessment across NCAR labs. These organization-wide data initiatives provide opportunities for staff to leverage knowledge, standards, systems, and lessons learned across labs enabling knowledge transfer among data professionals, engineers, and scientists.

4.1.4 Data Services

NCAR provides an array of scientific data services meeting needs throughout the research data lifecycle (see the Digital Curation Centre (2008) Curation Lifecycle Model as an example). Services are targeted at two levels: NCAR-wide and individual labs. This section presents a current snapshot of services for scientific data across NCAR plus a more detailed description of the three teams that I studied.

NCAR provides a variety of services for scientific data management and curation. Data sets and research products are available online in multiple locations such as the NCAR Community Data Portal, NCAR-Archive portal, Earth Observing Laboratory (EOL) data catalog, external repositories, and many individual lab and project web pages. Data collection and processing support are available in the science labs, usually provided by a data professional and/or by an EOL data manager for field campaign support. Metadata generation and quality services are provided by data professionals assigned to the science team; additional metadata support for archived data sets is available from NCAR-Archive and EOL. While Computing and Information Systems Laboratory (CISL) offers short-term data storage options to NCAR staff, long-term data archiving and preservation support reside in the two data archives, NCAR-Archive and EOL. Some science teams deposit their data and products in external repositories. Labs have internal staff for data analysis and engineering tailoring these services to their unique sub-disciplinary needs. Across NCAR, data services cover a range of Choudhury et al.'s service levels discussed in Chapter 2, supporting a range of research activities.

Looking more closely at the three teams involved in this research, this section reviews the data services offered. As previously mentioned, NCAR-Archive team started providing minimal data processing and archiving support back in the late 1960's. However, the services have

evolved to a full suite of world-class data services including preservation best practices, standardized metadata, searchable metadata database, web/cloud discovery and access, data citation, among others. These curation level services are targeted for the geoscience researchers, but the user community has grown to include a multi-disciplinary community of scientists, educators, and students.

As climate models have increased in complexity, growing expectations of data and model sharing in the climate sciences have propelled the standardization and sharing of products and attention to scientific data management. NCAR-Climate has a long history of offering the traditional services of data/model catalog, discovery, and access, but community interests have driven innovative service development in a shared, climate community data portal (e.g., Earth System Grid); data use guidance resources like Climate Data Guide; and a programming language for scientific processing and visualization, NCAR Command Language. The user community for these services is primarily the climate science community, but interviewees talked about increasing interest from other sciences and the general public. Many teams are funded by Department of Energy (DOE) and comply with requirements to archive data in a DOE repository.

Since the founding of NCAR-Solar in 1940, the lab has been providing data to the solar, heliophysics, and broader geosciences communities. Computing advancements have changed the data discovery and distribution methods, improving data transfer and time to make data available. Data services have evolved from the early days of manual data collection, storage on plates, and sharing in person to the current services of digital high-level data products, metadata generation, quality assessments, identifiers and citations, and sharing via web distribution with real-time access and preview movies. While the user community is largely the solar sciences,

high-level data products have allowed the broader earth science community to reuse these data sets.

4.1.5 External Drivers

The story of NCAR data services and efforts would not be complete without the mention of several key stakeholders and their interests. This section describes a few key stakeholders in the geoscience data community.

As previously mentioned, the National Science Foundation (NSF) is the primary funder of NCAR. NSF and other federal agencies and foundations have implemented data management and sharing requirements for funded research. In addition to funding agencies, prominent publishers and professional associations in the Earth Sciences such as the American Geophysical Union and American Meteorological Society have established data policies. The earth science community as a whole and sub-disciplines within it (e.g., climate modeling) have witnessed increasing expectations for data sharing. These requirements and policies have placed pressure on NCAR scientists to make data and research products available to the public.

Cyberinfrastructure projects are another growing type of stakeholder in the geoscience community. For instance, EarthCube is a geoscience cyberinfrastructure initiative to improve data sharing and reuse. In 2014, a meeting of the EarthCube Data Facilities Council, a multi-agency consortium, motivated the NCAR Deputy Director to prioritize data management needs and to support the establishment of the recent DSET initiative. These stakeholder groups and their interests have motivated NCAR to continuously innovate their data practices, knowledge, and services, and to make use of their flexible organizational structure to respond to the changing data landscape.

This profile provides an abridged historical and current snapshot of NCAR related to data

services and staffing. NCAR, as a research case, contributes how a national research center with a long history and strong scientific mission has developed and sustained data staffing and expertise. The next section is a timeline of key events related to NCAR data management and data expertise development, followed by the profile of Purdue University Libraries.

4.1.6 NCAR Timeline of Related Events

1940 High Altitude Observatory (NCAR-Solar) started

1960 NCAR was founded, absorbing High Altitude Observatory (NCAR-Solar)

1963 First supercomputer arrived

1965 Roy Jenne hired
Data Support Section (NCAR-Archive) started

1986 NSFNet started, NCAR was a member node

1987 Climate & Global Dynamics (NCAR-Climate) division formed with Warren Washington as first director

1990's NCAR-Archive worked on NCEP/NCAR reanalysis data sets

1993 NCAR-wide Internet and networking improvements ran until 1999

1995 NSFNet decommissioned; privatized Internet started
Information Infrastructure Technology Applications (IITA) started

2001 Data Management Strategic Plan created
Data Management Working Group (DMWG) formed

2003 New Data Service Manager of NCAR-Archive, name changed to Research Data Archive

2004 NCAR's first deposit into Earth System Grid

2007 Roy Jenne retired

2002 Community Data Portal launched
UCAR open publication & data policy announced

2011 National Science Foundation (NSF) data management plan required
NCAR library appointed first Research Data Scientist
Data citation working group (DCite) founded
Data management added in UCAR strategic plan

2012 American Geophysical Union's open data position statement issued
Climate Data Guide started

2013 White House memos on open data issued
NCAR stakeholder surveys conducted
American Meteorological Society (AMS) open data statement announced
New NCAR Director and Deputy Director appointed, supportive of data management

2014 EarthCube's Data Facilities Council first meeting
Data Stewardship and Engineering Team (DSET) formed
2015 AMS announced data archiving and citation recommendations

4.2 Purdue Profile

Purdue University Libraries, founded in 1876, is an academic library serving a multi-disciplinary campus with strong programs in engineering, agriculture, and business. In 2004, the library started exploring research data services as a new practice area. As previously mentioned, this research focused on two data teams and other professionals supporting data services in the library. This profile provides a summary of the organization, staffing, data efforts, services, external drivers, and timeline of key data-related events, focusing specifically on the library staff and also providing some organizational context for the study.

4.2.1 Organizational Overview

Innovation and *change* are terms prevalent in the history of Purdue University Libraries (hereafter referred to as Purdue). The library is driven by a mission to advance interdisciplinary learning, discovery, and knowledge. Located in West Lafayette, IN, USA, Purdue is an academic library serving Purdue University, a public, Big Ten university classified as very high in research activity (Carnegie Foundation for the Advancement of Teaching, 2015). The library was established seven years after the university was founded. See Table 4.3 for a summary of Purdue organizational characteristics containing primary activities, domain, and staffing and structure. Purdue has an extensive collection with over three million volumes and almost two million e-books and employs about 165 employees, including 89 professional librarians in 2014-2015 (Purdue University Libraries, 2016). The library is primarily funded by the campus but receives grants and contracts from Institute of Museum & Library Services (IMLS), National Institutes of Health, Andrew W. Mellon Foundation, Alfred P. Sloan Foundation, among others. As a

research library, Purdue aims to foster interdisciplinary research, knowledge sharing, and knowledge discovery among the campus. The library is reputed for innovative services and programs, receiving the Association of College and Research Libraries Excellence in Libraries Award in 2015. Purdue has been an early innovator in library research data management services and data literacy curriculum. With the arrival of the new Dean of Libraries, Purdue librarians started exploring the campus data management needs in 2004 and offering a suite of data services in 2007.

Short name	Purdue
Org Type	Academic library
Primary activities	Education and research support
Domain	Multidisciplinary, but strong campus programs in engineering, agriculture, and business
Funding	Campus and grants
Location	West Lafayette, IN
Founding year	1876
Staff size	About 165 librarians and staff
Divisions	• Archives and Special Collections • Health and Life Sciences • Humanities, Social Sciences, Education, and Business • Physical Sciences, Engineering and Technology • Collections Management • Circulation and Repositories • Resource Sharing • Resource Services • Instruction and Digital Program Services • Distributed Data Curation Center • Purdue University Press • University Copyright Office

Table 4.3. Overview of Purdue University Libraries

At the time of this writing, Purdue University, similar to peer universities, organizes campus units by academic domains and functions, centralizing similar expertise, professionals, and resources. One example is the campus IT unit encompassing high performance computing

and computer engineering staff and systems. The library follows a similar structure with several teams organized by functions and/or disciplines (e.g., archives, life sciences).

4.2.2 Data Staffing

Purdue includes two data-focused teams– Distributed Data Curation Center (Purdue-Archive) and Research Data Service team (Purdue-Consult) – and relies on a network of librarians across the library (Purdue-Lib). To leverage campus expertise and resources, the library partners on data services with campus IT and the sponsored research office, among others. Table 4.4 provides for a comparison of the library teams including primary activities, data staffing, services, and holdings.

Short name	**Purdue-Archive**	**Purdue-Consult**	**Purdue-Lib**
Full name	Distributed Data Curation Center	Research Data Service	N/A; other professionals across the library
Primary activities	Data repository services	Research data consultations	Varies by specialty
Funding	Campus, library, grants	Campus, library, grants	Library, grants
Founding year	2006	2011	1876
Primary data personnel	Data Curator, Software Engineer, Repository Outreach Specialist, Director	Data Specialists, Senior Data Specialist	Liaison Librarians, Metadata Librarian, Archivist, among others
Primary data services	• Data archiving • Data dissemination • Identifiers and citations	• Consultations • Training • Among others	• Consultations • Data reference • Metadata • Varies
Data holdings	Purdue data and related products	N/A	N/A

Table 4.4. Overview of Purdue teams

Research data services at Purdue started with librarians working in their respective library units. In 2004, the Interdisciplinary Research Librarian was appointed to explore the

collaborations for data services and the library's role in research data management, which is a position that no longer exists at the time of data collection. The dean realized that librarians needed an administrative title to be recognized as peers in data management by campus administrators, resulting in the creation of the Associate Dean for Research in 2005. Early data service initiatives included embedding librarians into research projects, allowing librarians to learn about everyday data practices and issues and scientists to benefit from the curation expertise of librarians knowledgeable about their project. Despite the successful initiatives, the Dean realized this model would not scale well if more researchers requested this service. A shift in service mode from one librarian to one project to one librarian to multiple projects happened where the focus became developing campus-wide services.

The new Associate Dean for Research founded the Distributed Data Curation Center in 2006 focused on optimizing research data management. The team included data professionals and the Interdisciplinary Research Librarian in the beginning. This center has been successful in receiving grants in the area of research data management and curation such as Data Curation Profile and DataBib, among others. With the growth of library data positions, a new team split off of the center, forming two library data teams: 1) data professionals providing consultations (Purdue-Consult) and 2) professionals providing repository services (Purdue-Archive). These teams are located in the Research Data Division in the library organization and receive support from liaisons, metadata specialists, and archivists located in other library units.

Purdue-Archive team investigates and pursues innovative solutions for curation--organizing, facilitating access to, and preserving research data. As NSF announced data management plans, Purdue University administrators and researchers were looking for campus solutions. Fortuitously, the library's e-Data Task Force (see Library Data Efforts for more

details) had explored data repository service and developed a prototype. Campus administration was impressed with the outputs of the e-Data Task Force and decided to fund the repository development. The library received campus funding and support to create a research data repository service, entitled Purdue University Research Repository (PURR). As previously stated, this team started as the Distributed Data Curation Center but has evolved to focus on a data repository solution, enabling archiving, preservation, and sharing of campus data and research products. This team has always employed a one-data-professional to many-researchers approach. At the time of data collection, the group included 4 positions including Data Curator, Software Engineer, Repository Outreach Specialist (vacant), and Repository Director. Given service demand and learning more about repository work, the Repository Outreach Specialist is a new position, splitting off from the Data Curator position, where the curator is more hands-on with the data and the specialist is more user- and community-focused. This team has received external funding to investigate and enhance the repository services and platform.

Purdue-Consult team is focused on providing advice, consultations, and designing new services to solve campus data management issues. When the Associate Dean for Research stepped down to a Senior Data Specialist position in 2013, library reorganization triggered the creation of a second data team focused on consultations and service design. The team consists of 3 Data Specialists and 1 Senior Data Specialist. The data professionals have different backgrounds and areas of expertise (e.g., big data, social science, metadata), allowing them to serve a broad range of disciplines and data needs. This team uses a one-data-professional to many-researchers strategy. Similar to Purdue-Archive, this group has received grants and contracts to develop innovative programs, tools, and services.

Purdue-Lib is the final library player comprised of a network of librarians and professionals across the library. A few examples of positions include Liaison Librarians, Metadata Specialists, and Archivists. These positions are housed in several units across the library. All library professionals bring their expertise to support or participate in research data services. For instance, the liaison librarians are a service access point, where they refer researchers to the Purdue-Consult or -Archive teams. In the early years of data services, liaisons were hesitant to be involved but, over time, liaisons have learned more about research data and grown more comfortable with their role in data services. A few liaison librarians have even gone beyond their duties to design data curriculum and serve on data-related projects for their designated communities. In addition to liaisons, the Archivist extends her/his extensive knowledge of preservation strategies and standards to the repository practices and staff. S/he has been an active participant in designing and standardizing archival workflows in the development of PURR and continues to consult on data preservation issues with the data professionals as needed. Historically, these librarians did not have research data as a formal responsibility but collaborated to support the library's goal of data services. However, recent liaison position descriptions have included research data management as a duty.

The Dean of Libraries has been resourceful and opportunistic in funding new data positions. Data services gained new positions from the retirement of reference and cataloging librarians, the campus and grant funding for data projects, and campus faculty cluster hires in big data, systems biology, and plant sciences. These efforts were able to fund positions for the data-focused teams and new metadata specialist and liaison librarians that were targeted toward research data.

My study explored data teams and personnel in the Purdue Libraries. However, additional data management professionals, teams, and services may exist across campus in other academic departments or research centers.

4.2.3 Library Data Efforts

Three data–related working groups emerged in the Purdue Libraries related to research data services. The first group was the e-Data Task Force started in 2008 to define the data repository service and make recommendations on data collections, policies, and practices. A key outcome of this group was a report outlining recommendations for policies, staffing, infrastructure, and sustainable funding to extend library services to a data repository. This task force produced a service model illustrating service features, users, and specific librarian roles as well as developed a repository prototype using Fedora and testing it with 6 data sets. The NSF data management plan announcement in 2011 advanced the data repository service to a campus priority. The work of this task force provided a campus data solution and received campus funding for the PURR development and implementation.

The PURR working group was established in 2011 comprising liaison librarians, archivists, library data service staff, and campus software developers. The initial goals were to investigate and determine whether the HubZero platform for scientific collaboration, developed at Purdue University, would work for research data curation. This group focused on technical requirements, policies, and librarian roles. A set of recommendations and a preservation policy were key products from this working group. This working group developed, adopted, and implemented a successful data repository service for the campus, while offering an exploratory project for librarians to learn about data archiving.

In response to the e-Data Task Force recommendation of liaison librarians' role in research data services, liaison librarians founded the Data Education working group in 2011 to provide staff training on data management and curation. Key products of this group were workshops and a libguide for librarians on data management resources. This task force offered opportunities for librarians to learn about data curation and their new roles in data services.

These three library efforts spurred the library staff to explore the research data landscape fostering individual learning and service enhancements.

4.2.4 Data Services

Since the official inception of data services in 2007, Purdue has evolved and grown its services and programs to respond to the research data management needs of the campus. In the early years, data librarians were embedded in research teams to consult and advise on data practices. The library data professionals realized the embedded data professional approach would not scale, resulting in a service direction change. The library designed a suite of services and trainings based on insights from these early experiences. The current services include consultations on data management planning, metadata, big data, or other issues; data management and literacy trainings targeted at different audiences; data repository; persistent identifiers; data citations; data reference; among others. Data professionals are continuously identifying new trends or needs and brainstorming new services, resources, tools, or programs to assist Purdue researchers in data management and curation.

A key service is the Purdue University Research Repository (PURR) offering a collaborative workspace and data archiving and sharing platform. Started in 2012, PURR runs an instance of HubZero, a scientific collaboration platform created by researchers at Purdue University, providing a suite of collaboration tools (e.g., wiki, messaging, virtual machines) and

data publishing tools (e.g., metadata, archiving, sharing). Purdue researchers are allotted 10 GB for each project for three years and 1 GB of published data at no cost. Published data in PURR are maintained for at least 10 years and at the end of 10 years transferred to Purdue special collections. More storage is available for purchase. The Purdue-Archive staff is available to assist researchers with curating data sets. This level of data services aligns with the preservation level in the Choudhury et al. typology.

All the data services are targeted at the large community of Purdue University faculty, researchers, and students. Purdue data professionals and librarians are offering a variety of services tailored to their constituencies.

4.2.5 External Drivers

Similar to the NCAR story, external drivers and stakeholders motivated Purdue data services. Starting as early as 2008, campus administrators and a few researchers began hearing about NSF proposed requirements for data management plans. The National Science Foundation is the largest sponsor of research at Purdue University. Campus department heads and research administrators realized researchers needed help with data management and were looking for solutions. Since 2011, federal funding agencies, foundations, the White House, prominent publishers, and professional associations have established a series of data requirements and policies. External drivers have prioritized research data management at the campus level, resulting in support and resources for data services and infrastructure.

The next section is a Purdue timeline of events related to research data services.

4.2.6 Purdue Timeline of Related Events

2004 New Dean of Libraries arrived
Librarians explore role in data management
Interdisciplinary Research Librarian position created

2005 Associate Dean for Research position created

2006 Purdue Libraries launched institutional repository for documents
Distributed Data Curation Center started

2007 Data services started
Data curation profile grant awarded, Purdue University Libraries as a partner
Library restructuring

2008 e-Data Task Force started

2011 Purdue University Research Repository (PURR) working group started
Data Education working group formed
DataBib grant awarded
Data literacy grant awarded
Library restructuring
National Science Foundation (NSF) data management plan required

2012 Purdue University Research Repository launched

2013 White House memos on open data

2015 Library data teams all relocated to one office in prominent campus location

This profile provides a summary of Purdue University Libraries related to research data services and staffing. As a research case, Purdue offers a unique investigation of an academic library with an innovative culture and experiences in developing and supporting research data services for a multi-disciplinary community.

This chapter provided a description of the two sites in order to contextualize the study results in Chapters 5 and 6. The next chapter focuses on the data teams, staffing, roles, and expertise that emerge from the process of building research data services.

CHAPTER 5: RESEARCH DATA STAFFING AND EXPERTISE

The trends in research data management are placing new demands on researchers and research institutions to publish research data sets. Data centers like NCAR and academic libraries like Purdue are responding to these changes by building internal support and services for data management and curation. The research for this dissertation investigated the data staffing and expertise that emerge from the process of developing data services at the two research sites, highlighting differences by context. The analysis examined what staff roles and expertise were involved in research data management services.

A key product from the analysis is the presentation of two models for supporting data management and a salient set of elements in these models. The two research sites employed different models for embedding research data expertise in the organization, in terms of organizational structure, position design and roles, and categories and configurations of data expertise. The set of data roles aligned with the expertise categories in definition and in function.

For this study, the term, research data expertise, signifies the type of knowledge, skills, and experience needed for data work, including the social dimensions learned by performing the work. The chapter organizes results into two sections, 1) data staffing approaches and 2) research data expertise, concluding with a summary of key findings.

5.1 Data Staffing Approaches

My study documented two models for building research data expertise and services. The models of NCAR and Purdue shared a set of common elements for supporting research data services that emerged from the cross-case analysis: organizational structure, boundary-spanning data positions, roles and needed expertise. See Table 5.1 for a comparative summary of these models.

Elements	NCAR model	Purdue model
Organizational structure	Data-focused teams Research teams Service partners: External data archive used by some projects	Data-focused teams Service partners: Library staff (e.g., liaison librarians, metadata specialist) Campus staff (e.g., IT, research office)
Positions	Generalist, spanning several roles	Specialist, comprised of one or two roles
Data roles	Data liaison, Data curator, Data engineer, Data scientist, Data service manager	Data liaison, Data curator, Data engineer, Data service manager
Expertise categories emphasized	Data, Research, Curation, Engineering, Service, Analytics, Leadership	Data, Research, Curation, Service, Leadership

Table 5.1 Summary of NCAR and Purdue models

This study elucidated a set of common elements in data service models – organizational structure, positions and roles, and expertise categories and configuration. In terms of data service staffing, data professionals were placed in two types of teams: 1) science teams working beside the scientists in labs and 2) data-focused teams serving researchers in other units. The design of data positions evolved over time at both sites, combining different data roles and adjusting the generalist vs. specialist nature of the position. I identified a set of 5 data roles - Data liaison, Data curator, Data engineer, Data scientist, and Data service manager. These roles highlight how data work is conceptualized and organized into these organizations. The summary of the organizational structure for data services is followed by how the data positions are designed to support research data services.

5.1.1 Organizational Structure for Data Services

The two sites used different team structures to support research data management and curation. The models included common elements of teams and dynamic collaborations across teams and employees. To illustrate these findings, I describe each site's staffing approach concluding with general observations on these team structures.

NCAR structure

The NCAR model placed data professionals in both data-focused teams and science teams. Data-focused teams involved a group of data professionals providing data management and curation services but located in a separate unit from the science labs. The NCAR-Archive team embodied this data-focused team structure. NCAR-Archive was a team of data professionals, housed in CISL, a computing-focused laboratory. This team served all the NCAR science labs providing expertise, practices and systems for scientific data/model preservation and sharing. The team structure included 8 Data Engineer positions that all provided curatorial and engineering support – metadata, preservation, training, and discovery – with the Data Service Manager providing coordination and direction. As will be described later, the team had extensive knowledge of data preservation standards, best practices, and technologies as well as collection development for archival data sets. Given that this team served both users of observational data and simulations, staff members had a broad knowledge of geoscience research topics, data types, and analysis techniques. While the primary activity of the NCAR-Archive team was data archiving service, several data professionals in other labs drew attention to this team as an internal resource for data archiving and preservation expertise and guidance.

A science team approach placed a data professional into a research team to offer on-site support for data management and curation. Both NCAR-Solar and NCAR-Climate utilize these science teams, where data professionals work side-by-side with scientists and other research staff on solving data challenges. In this approach, data professionals were assigned to an instrument or simulation project(s) with responsibilities for data collection, management, description, and dissemination. These data professionals worked across the multiple stages of the research data lifecycle, fulfilling a generalist role. The placement of data professionals into science teams put

these experts beside scientists in the field or in the lab, offering their data expertise as data challenges arose. Similar to the NCAR-Archive team, these professionals provided curatorial and technical expertise to support data management with a few nuances. At NCAR-Climate, the climate and weather modeling nature of the work motivated their Data Engineers and Data Scientists to be knowledgeable about the different model types and their features. Since NCAR-Solar lab aimed to create valuable, high-level data products, all the data professionals needed an understanding of how end-users would work with and analyze these data. This science team approach offered support for day-to-day data management needs allowing these labs to cultivate the expertise areas to meet the needs of their sub-discipline or research topic.

Many of the NCAR team configurations represented a historical artifact of the lab organization, values, and priorities. Early in the history of NCAR-Archive management set priorities for data sharing and hiring staff with the expertise to support data preservation and dissemination. The first manager prioritized hiring staff that had experience working with different data types and atmospheric sub-disciplines, while the second manager continued this staffing emphasis but added a new priority of bringing technical and outreach expertise to the NCAR-Archive team. While my study focused on teams with data services, NCAR interviewees noted that some science teams did not hire data professionals and did not place a high value on data management and sharing. NCAR labs had tremendous autonomy to configure teams and positions to meet their priorities and needs, independently of how other labs were designing teams and services.

The NCAR model of placing data experts into both science teams in concert with data-focused teams balanced the particular needs of each science lab and provided consistency across the organization. Science labs benefited from local data experts that understood the research

techniques, instruments and/or models, data types, and cultural norms specific to their sub-discipline. The labs also were able to draw on a deep expertise in geoscience data preservation and sharing from the NCAR-Archive team, serving users across NCAR. This model situated data professionals throughout NCAR, not always uniformly distributed, to support cutting-edge atmospheric data management.

Among the supplemental data centers, the most prominent trend was the placement of data professionals into data teams, occurring at 9 supplemental sites. These sites were compromised predominantly of academic and government-funded research centers. The NCAR trend of data professional placement into both data and science teams was corroborated at 4 data centers; these sites were large, national research centers with multiple data archives. A final trend in the supplemental data centers (5) was the placement of data professionals only into science teams. The formation of science and/or data teams to support data management operations was a prominent type of service model across the supplemental data centers, demonstrating the transferability of this finding to other centers.

Purdue structure

Purdue placed data professionals into data-focused teams supported by a network of service partners providing expertise for research data services. For research data services, Purdue honed certain data expertise locally but relied on existing expertise across campus. This network approach allowed Purdue to leverage existing expertise and systems in other units. As the Repository Director described: “Our approach has been to partner: other people have those expertise and infrastructure. They can do much better than we can” (Purdue 101). Since the inception of data services, the library cultivated two data-focused teams and collaborations with professionals within and outside of the library.

The Purdue-Archive team was a data-focused team responsible for the data repository and sharing services for the campus. This team provided an expertise in archiving, preservation, and sharing of multi-disciplinary research data and products that does not exist anywhere else on campus. Similar to the NCAR-Archive team, the 4 data professionals served the data archiving needs of the entire university. The team included the positions of Data Curator, Software Engineer, Repository Outreach Specialist, and Repository Director. In comparison to the NCAR data professionals, these data professionals held more specialist positions, where each position was responsible for different aspects of repository work (e.g., curation, engineering) as their titles imply. As will be discussed later, the team members brought extensive knowledge of data preservation, user interface design, relationship building, and service design to their work. Although these professionals were housed in the library, the team was available to researchers across campus, providing services to the whole Purdue research community.

The second data team, Purdue-Consult, focused on data consultations, training, and designing services addressing the needs of Purdue researchers. The group included three Data Specialists with a Senior Data Specialist providing leadership and coordination. These data professionals exhibited deep knowledge of cross-disciplinary research practices, service design, big data, and metadata standards. For instance, one Data Specialist handled any requests on big data due to his/her research background, and a second Data Specialist focused on metadata questions given her/his extensive knowledge of scientific description and standards. This team offered data management guidance, resources, best practices, and trainings to the entire campus.

While the two data teams in the library provided data expertise to the campus, the teams were supported by a network of service partners inside and outside the library for additional knowledge and resources. Data professionals received support from liaison librarians, metadata

specialists, archivists, IT specialists, software engineers, and research compliance specialists. For instance, the liaison librarians brought extensive knowledge of disciplinary standards, norms, and practices of their designated communities, and IT specialists provided expertise in supercomputing techniques and infrastructure. While many of these partners did not identify as data professionals, they could extend their professional expertise to research data enhancing the library data services and expertise. The network approach allowed Purdue to cultivate local knowledge, professionals, and practices while utilizing existing expertise available on campus.

The Purdue data staffing approach was a learning artifact reflecting their insights about data management practices and community needs. Initially, Purdue implemented an embedded approach where data librarians were placed into research teams to support data management and identify data service opportunities for the library. While this approach was successful in promoting good data practices and understanding individual researcher's data practices, the Dean realized this approach would not scale to meet the growing campus needs. Based on their experiences, the library created a suite of data consulting, management, and preservations services supported by two teams of data professionals plus a network of partners, bringing together a variety of professionals, expertise areas, and resources to address campus data needs. The library has autonomy to configure and reconfigure the teams and positions to meet their service priorities and campus needs. With data professional turnover, the library has established a reflective process where each position is reviewed for alignment with library goals and current services needs.

I observed variation in the data staffing models in the set of supplemental interviews from academic libraries. Four libraries employed teams dedicated exclusively to research data services similar to the Purdue model. Other organizational approaches included: solo librarian without

additional support (1), solo librarian relying on a network of partners (3), and multi-functional library teams serving data and other functions (4). Two libraries were in the early stages of planning for data services, and their staffing models were too nascent to classify. To meet the campus data needs, libraries commonly partnered with campus units; frequently mentioned partners included the campus units of computing, sponsored research, and the institutional review boards for research with human subjects.

Cross-case Summary of Structure

The study contributed two models for organizational structures to staffing research data services. The data-focused teams located data professionals in a separate unit from researchers providing data services as requested. In contrast, the science team approach co-located the data professional and science staff providing day-to-day support for data management and working beside researchers in the field, lab, or office. The proximity of data professionals to the science work was a difference observed in the type of teams and their placement in the organizational structure. The second theme was the dynamic interactions and collaborations between professionals in the organization. The interactions involved two types – 1) data professionals providing data services to scientists, and 2) data professionals collaborating with other professionals in the organization to offer data services. See Figure 5.1 for a visualization of the models at NCAR and Purdue.

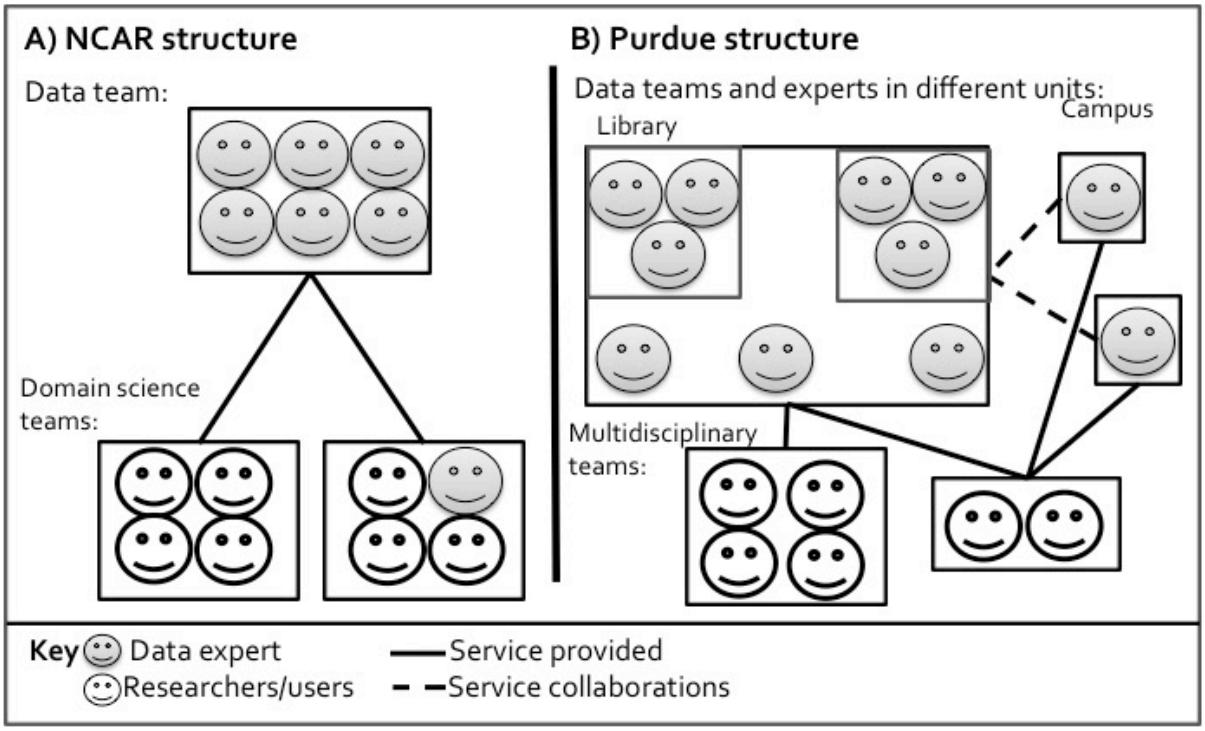

Figure 5.1. Team structure for data professionals

The two panels in Figure 5.1 summarize the two models supporting data management in the research sites. NCAR (see panel 5.1a) placed data experts in both data-focused teams and in science teams, whereas Purdue (see panel 5.1b) depicts two data-dedicated teams with a network of service partners located throughout the organization available for research data management support. The study documented two models of team structures and service collaborations that were effective in supporting these research communities. The next section continues the theme of data staffing by highlighting data positions and roles that emerged at the two research sites.

5.1.2 Boundary Spanning Positions

As NCAR and Purdue developed their data services, new positions and roles for data professionals have appeared and evolved in the organizational structure. Data roles are the categorization of the work activities and professional areas of practice of particular employees. These roles emerged from the interview data where participants described their work activities

and duties. I prioritized the participant descriptions over the descriptions in job advertisements since data positions were not well represented in the job classification systems at the research sites.

Data professionals were boundary spanners (Friedman & Podolny, 1992; Tushman, 1977), bridging people from different disciplines or professions (e.g., earth sciences, social sciences, engineering, information) to optimize research data management, sharing, and reuse. The boundary spanning nature of data work resulted in data positions comprised of multiple roles (e.g., data engineer, curator, and data scientist). Some non-data positions contained a data role (e.g., solar scientist responsible for data management). See Table 5.2 for the description of data roles identified at the two research sites.

Roles	**Definition**
Data curators/ managers	Responsible for the curation of research data; primarily concerned with ensuring use and access.
Data engineers	Building software, automating processes, and computational solutions for data.
Data liaisons/ consultants	Provide guidance, best practices, and resources for users via consultation, instruction, etc.
Data scientists	Designing simulations and high-level data products; responsible for data management and quality.
Data service managers	Oversight for data services/units; accountable for service activities and performance.

Table 5.2. Composite of data roles for both sites

The curator roles ensured the quality, access, and use of data throughout the research lifecycle. While some curators were responsible only for certain lifecycle stages (e.g., archiving or processing), others worked across the lifecycle and its activities. The curator roles were supported by the data engineer roles that design and maintain tools, applications, and systems for research data. The data scientist role was distinguished from the curator by contributing to analysis and visualizations and creating complex, high-level data products for the user

communities. The liaison or consultant role supported data producers and users in their research endeavors by offering advice, best practices, and trainings to enable quality data management. Finally, the data service manager role provided the vision and motivation as well as oversight and coordination of data work activities. Data positions were designed to combine more than one role such as data curator and engineer or data liaison and curator, and the case results included several position configurations. These roles illustrate how organizations can conceptualize data work and responsibilities, structure positions, and embed research data expertise into teams or units. Case and cross-case themes that emerge from my study data on positions and roles for data management follow in the next section.

NCAR data positions and roles

NCAR had multiple data positions and team configurations for scientific data management and curation. The roles for data management emerged to meet each lab's data needs. All roles listed in Table 5.2 were discussed by the NCAR interview respondents. This section summarizes the data positions and roles of the three teams that I studied.

NCAR-Archive is a team of data professionals bridging atmospheric science, archiving, and engineering areas of practice and modifying data positions to keep pace with data trends. Currently, the team is comprised of 8 Data Engineers led by a Data Service Manager and a Senior Data Specialist. The staff roles have evolved over the years. In the beginning, the team began with two staff members performing curation work:

> However, with hundreds of millions of observations, it is clear that the amount of manual intervention involved in the cleanup process must be limited. Thus there will always be some problems in the various data sets; however they are usually reduced to the point that the data can be easily used. (Jenne, 1975, p. 2)

The deliberate adoption of more technology into data work started in 2003 when the current Data Service Manager assumed leadership. The data work incorporated more engineering tasks, than under the previous manager, to utilize databases, code libraries, programming languages, networks, and web technologies to automate processes and improve distribution. Under the current manager, the emphasis on service activities grew and data professionals were responsible for responding to user requests, outreach, and education. Data Engineer positions have evolved into a blend of Data Curator, Consultant, and Engineer roles. All data professionals conducted data quality checks, metadata generation, and preservation techniques, as well as responded to user requests and developed software to expand the features of the data archiving system. The Data Service Manager position has the additional role of Service Manager added to his/her position, providing leadership and coordination. NCAR-Archive has evolved data positions and roles to meet the changing demands of their user community.

At NCAR-Solar and Climate science labs, the data roles and their evolution exhibited similar trends to NCAR-Archive. The curation role of these positions has grown in importance with the data and model sharing expectations of the geoscience and broader community. The engineering work of data professionals also has increased with the advancement and adoption of technology in these labs. Data professionals assumed a combination of roles to meet the needs of their assigned project. Of the six data professionals interviewed, two configurations of roles emerged across the labs. Three positions included the combinations of Data Scientist, Curator, and Consultant roles, while the other three positions comprised roles of Data Curator, Engineer, and Consultant. Multiple roles combined in one position offered the science teams a professional that could bridge many professional boundaries and solve a variety of data challenges.

Despite the history of data management efforts at NCAR, data positions were not

formally recognized in the job classification system – interviewees held positions of Project Scientist or Software Engineer with working titles including the term, *data*. As one NCAR-Climate Data Engineer summarized the situation: "I mean people definitely recognize what I do is important…I'm filling a role that they don't quite understand and that they don't have a slot for me" (NCAR 206). To design data positions, managers have to modify Software Engineer or Project Scientist positions to meet the needs of data management work. Another theme from the interview data was the lack of a career ladder for data management at NCAR. Several data professionals expressed concerns that data positions moved them off the traditional scientist or engineer career tracks. The Human Resources unit at UCAR manages the NCAR job classification and matrix system. As true in most organizations, Human Resources processes have an impact on the ability to design positions and career advancement opportunities that reflect the changing nature of data work and roles.

The data and boundary spanning roles observed in the NCAR data positions were evident across the positions at the supplemental data centers. The roles of data curator, scientist, engineer, consultant, and service manager were observed at all 18 data centers but were combined in unique ways to meet the organizational needs. A prominent trend in data positions construction was a single position combining the roles of data engineer, curator, and consultant, observed at 13 supplemental data centers. A second popular trend was a data position comprising the roles of data scientist, curator, and engineer, utilized at 10 data centers. Similar to the NCAR case, the supplemental data centers were constructing data positions to combine multiple data roles, crossing professional boundaries to address data issues.

Participants at other data centers expressed similar challenges in establishing data management career ladders and job classification systems. Of the 18 data centers, none reported

career ladders specific to data management. However, 13 centers utilized established, non-data job tracks (e.g., management, engineering, science) to provide advancement opportunities for data professionals. This workaround method was problematic for data professionals as career advancement meant the reduction in the range of curatorial responsibilities.

Purdue data positions and roles

As the library learned about data management, the roles for research data changed to meet the needs of the Purdue campus and to complement the existing infrastructure. In the early years of data services, the library placed the Interdisciplinary Research Librarian and Liaison Librarians in research projects for hands-on curation, consultations, and service design while the Associate Dean for Research liaised with campus administrators on service development and collaboration. Data roles in the early years focused on Data Consultant and Curator. The development of a data repository brought a new need for technical skills resulting in a role of Data Engineer added to the library data staff. Finally, two librarians involved early in data services advanced to leaders of the data-focused teams, serving as the role of Data Service Manager. At Purdue, library data positions evolved to include the roles of Curator, Engineer, Consultant, and Service Manager; the Data Scientist role was absent in the Purdue staff. In contrast to NCAR, Purdue has distributed the role of data liaison across the library to liaison librarians and other library staff serving as the first service access point. Refer to Table 5.2 for role descriptions.

The Purdue-Consult team has established a series of data positions with similar role configurations. As already noted, the Purdue-Consult team was comprised of 3 Data Specialists led by a Senior Data Specialist. All data specialist positions combine the roles of Data Consultant and Curator; the senior specialist has the additional role of Service Manager. Since its inception

in 2011, this team utilized similar position/role design for its data positions. The combination of Curator and Consultant roles enable Data Specialists to leverage the library expertise in liaison and curation work. These data professionals are collaborating with library, research, and computing professionals to offer responsive data services to the Purdue campus.

The Purdue-Archive team included a variety of positions with different role combinations; these roles emphasized the specific expertise that each staff member brought to repository work. The current positions are Data Curator, Software Engineer, Repository Outreach Specialist, and Repository Director. Over time, this team has added positions and evolved the positions and roles. For instance, the initial repository specialist position, comprised of Data Curator and Consultant roles, was a primary service provider for the campus. This position has split into two positions each emphasizing a different role. The Repository Outreach Specialist has the primary role of Consultant and liaising with the community and secondary role of Curator, while the Data Curator focuses on the Curator role first and Consultant role is second. As the repository director explained, the library had a nascent understanding of data librarian positions in the beginning: "…we know what a cataloguer does but this [data librarian] isn't the same thing. A lot of the same principles but applied to a different space and that has been a huge challenge too. And that is reflected in the reorganizations" (Purdue 101). The data positions and teams in the library have evolved as the services were defined and demand grew.

Since the beginning of data services in 2007, staff turnover has allowed data positions to be re-envisioned and re-structured to meet the evolving campus needs. Purdue data service positions have evolved to include some tenure-track librarian positions, setting them apart from many of their peer institutions. The quote illuminates the recent trend of tenure-track positions for research data services: "when data [services] first started here, it was soft money positions. It

wasn't even a permanent tenure track position and, to this day, only one person has tenure who has a data position…nobody else has because it's so new" (Purdue 104). The creation and refinement of data role and position design continued to evolve at Purdue.

I observed similar trends in data roles for research data services in the supplemental library interviews. A popular approach seen in 9 libraries was the data position combining data curator and liaison roles. The trend of combining 3 data roles into a position was less common in the set of supplemental libraries. Two libraries had a data position combining data curator, liaison, and engineer roles, while 3 libraries had 1 position comprised of data curator, liaison, and manager roles. Five libraries had data positions focusing primarily on the consultant role. One library was adding a new position including data science responsibilities. The supplemental sites confirm that all data responsibilities appear to be covered by the 5 roles, but that the arrangement of those roles into data positions varied in the library community.

5.1.3 Specialist vs. Generalist Position Trends

The cross-case analysis on staffing revealed differences in the generalist vs. specialist trend of data positions. At NCAR, the trends in data positions moved from specialist to generalist, where data professionals perform a larger variety of data activities, such as processing, metadata, archiving, and consultations, than in previous years. In contrast, specialist data positions were a popular trend in the supplemental data centers. Thirteen data centers had specialist positions such as Metadata Specialist, Archivist, Data Scientist, etc.; however, 6 centers had generalist positions for data management. Data professionals at both NCAR and data centers foresaw that the specialist positions for data management would evolve to generalists, where data professionals assume many roles, due to technology advancement and/or funding limitations. Three center participants noted the increase in generalist responsibilities for those

previously in specialist roles posed a big educational challenge.

Specialist data positions are a predominant trend at Purdue. The data positions at Purdue-Archive transitioned from generalist to specialist positions, where the growth of data services spurred the increase in data positions and refinement of roles and positions. The Purdue-Consult team included a set of specialist data positions. The set of supplemental libraries exhibited a variety of generalist and specialist positions for data management. In particular, the generalist data librarians required a broader knowledge of all activities in the data lifecycle, practices of multiple disciplines, and technologies.

5.1.4 Summary of Data Staffing

My study has documented the organizational structure supporting data services highlighting a set of 5 data roles and trends in boundary spanning roles and specialist vs. generalist data positions. With a strong scientific mission, NCAR placed data professionals beside scientists in the lab providing day-to-day support, and in data-focused teams, serving the data archiving needs of all the labs. In contrast, the organizational context of Purdue University enabled the library to cultivate local expertise among their staff but rely on campus partners for deeper knowledge and competence in other areas. These two promising approaches for teams and partnerships allowed the organizations to provide data expertise and effective data management solutions. The evolution of data position design was observed in both research sites. Data professionals work across many professional and disciplinary boundaries to provide effective data management and curation services. The boundary-spanning nature of data work resulted in data positions with multiple roles – Data Curator, Engineer, Scientist, Consultant, and/or Service Manager. While the trends of generalist and specialist nature of data positions differed by the two sites, the set of 5 data roles in Table 5.2 represents the responsibilities of data

professionals and ways they contribute to the scientific enterprise.

5.2 Research Data Expertise

As the data services and staffing evolved, a set of knowledge and skills requirements for data professionals emerged to meet the organizational needs for research data management. While similar expertise themes developed at both sites, the emphasis on certain areas or configurations of expertise varied slightly between the sites. This chapter section begins with the presentation of research data expertise types and moves to the expertise configurations in data professionals.

5.2.1 Research Data Expertise Types and Categories

Handling and preserving research data required data professionals to have a multi-dimensional expertise varying in areas and their depth. Eighteen distinct types of research data expertise were identified through iterative coding related to the knowledge areas and work performed by data professionals. A description of the expertise type and category construction process was included in section 3.6.2. Table 5.3 provides definitions for each type of expertise. The definitions are a composite of participants' descriptions across the research sites. In interviews, participants emphasized certain expertise categories over others. The table presents categories in descending order of emphasis by NCAR participants, starting with strongly emphasized categories.

Research data expertise types are associated with seven categories of functional areas – data, research, curation, engineering, service, analytics, and leadership. Work in the data and research categories requires an understanding of the research process, techniques, and data types as well as the landscape of data trends relevant to data work. The curation category encompasses the core skills of organizing, preserving, and standardizing data and products. The curation

category is supported by the engineering category, creating and customizing technologies, software, and systems for research data management. The service category represents the importance of relationships with users and collaborations in offering data services. These categories highlight the valuable contributions of data professionals within their institutions and areas where organizations can focus staff development and learning efforts. The following section provides only cross-case details on each expertise category; I have included the detailed case results on expertise types and categories in Appendix D.

Categories	Types of Expertise	Definition
Data	1. Data handling	Common techniques for data processing, wrangling, and manipulating.
	2. Data landscape	Data trends, stakeholders, mandates, and issues.
Research	3. Research process	Research-data lifecycle, workflows, and activities.
	4. Research instruments or models	Common research models or instruments used and their differences.
Curation	5. Organization	Organization of research data and products into collections including metadata and retrieval.
	6. Standardization	Data-related standards, mandates, and requirements; and developing compliant products and practices.
	7. Preservation	Preservation planning, strategies, and technologies; provenance; and translating archival best practices to research data.
	8. Data quality	Quality assessment of data and/or metadata, best practices for quality products.
	9. Ethics	Ethical and legal issues related to research data such as privacy, intellectual property rights, and licensing.
Engineering	10. Engineering	Software engineering and customizing existing software and systems to solve data problems.
Service	11. Data uses & users	User needs, potential uses of data, and user-friendly service design.
	12. Data discovery	Data reference interviews and accessing licensed data.
	13. Training	Instructional design and evaluation targeted at researchers, users, and organizational staff.
	14. Relationship-building	Trust-building with users and communication skills.
	15. Collaboration	Team player, and understanding of stakeholders and their interests.
	16. Data metrics	Archive performance & data value assessments, including collecting measurements & calculating metrics.
Analytics	17. Data analysis	Analytical or visualization techniques commonly used.
Leadership	18. Leadership	Management theory and best practices in designing teams and work, and service planning and implementation.

Table 5.3. Compilation of research data expertise types and categories

Data

The data category refers to competence in working with data including common techniques for handling, transforming, or manipulating data as well as familiarity with commonly used data types, structures, and formats. Data work also involved staying current on the trends in the data landscape. While both of the research sites strongly emphasized data skills and trends, the scale of expertise development was different. The NCAR data staff had a deeper knowledge of data, techniques, and trends in the geosciences, than staff at Purdue. All NCAR data workers had previous experience working with atmospheric or related geoscience data. A few NCAR data professionals had extensive experience in managing geoscience data. These types of expertise provided a solid foundation for data professionals to support research endeavors.

Research

The Research category highlighted the intimate knowledge of the research process, workflows, and activities that data professionals bring to the work. NCAR data professionals only emphasized the importance of knowledge of scientific instruments and models as well as understanding the history of these techniques. Similar to the Data types of expertise, the scales of expertise within the Research types were different between the two sites. NCAR has honed a deep level of expertise in research processes and instruments that is specific to their domain user community, whereas Purdue has cultivated this type of expertise to a certain point in their staff and relied on service partners to provide more knowledge and competence in research workflows and instruments.

Curation

Data professionals at both sites were involved in activities to further cultivate their curatorial knowledge and competence. This category included five expertise types revolving around information organization, preservation, data quality, standardization, and ethics. These types of expertise were comprised of traditional library and information science skills. Both research sites valued the curatorial expertise, but the length of time cultivating this expertise differed by the sites. At NCAR, curatorial expertise has been a recent priority for data professionals to develop. Purdue has a long history of staff with LIS expertise and focused recent efforts on extending this curation expertise to research data objects. This category aligns with the staff role of Data Curator, and data professionals that assumed this role exhibited deeper levels of curation expertise.

Engineering

The Engineering category highlighted the ability to harness technology for data solutions. While engineering expertise emerged at both of the research sites, the scale of expertise development was different. The NCAR data staff had a deeper knowledge of programming, software engineering, scientific technologies, and computing trends than staff at Purdue. Most NCAR data professionals were comfortable working with several software, operating systems, and programming languages. Engineering expertise had great alignment with the Data Engineer role and some alignment with the Data Scientist role, where these professionals needed strong expertise in this area to perform the work.

Service

The research sites had similar themes in relation to service expertise. The Service expertise category encompassed skills for designing and offering services to skills for engaging users and stakeholders. While these skills were important in both sites, Purdue librarians placed a

stronger emphasis on this type of expertise than NCAR data professionals did. The service expertise category aligned with the role of Data Liaison/Consultant, emphasizing the importance of the service function.

Analytics

The analytics category drew attention to the knowledge and experience with data analysis and visualization techniques. This expertise type was present only in the NCAR data. Analytics expertise emerged as an important skill set for NCAR data professionals. While analytics was not an expertise developed in the library staff, the presence of analytics knowledge in other campus units allowed the library to leverage existing domain-specific expertise in analysis techniques. The Data Scientist role aligned with these categories, drawing attention to the analysis and visualization side of data work.

Leadership

Data services required the ability to plan and coordinate several tasks and actors (e.g., data professionals, scientists, service partners, and stakeholders). Leadership was not emphasized by either NCAR or Purdue participants as an important skill set for data work. However, the expertise in leadership and management enabled the design and execution of research data services. The Leadership category of expertise had great alignment with the Data Service Manager role in function.

This section summarized the cross-case findings on emergent research data expertise types and categories that represent the knowledge and skill requirements for data professionals. The 18 types of expertise comprised 7 related functional areas of data, research, curation, engineering, service, analytics, and leadership. Appendix D provides further details of the individual case results for each expertise type and category.

The supplemental sites confirmed the research data expertise types. The participants at supplemental data centers mentioned all the types of expertise but the emphasis on the importance varied from NCAR results. Overall, the data centers placed less emphasis on the ethics, training, collaboration, and data discovery types. A few data centers highly valued staff with deep expertise with data movement and storage for large-scale data. I observed similar trends in the expertise types between Purdue and the supplemental library interviews. The expertise types of data handling, data landscape, and research process were strongly valued by the library participants and identified as areas that they were trying to cultivate, validating this finding in the Purdue case. The types of information organization, standardization, data use and users, training, relationship-building, preservation, collaboration, and data discovery were reported as required skills for data librarians. No library participants mentioned the expertise type of research instruments and models.

The next section follows this theme of expertise by illustrating the combination and levels of expertise of data professionals.

5.2.2 Configuration of Expertise

A second theme from the analysis on research data expertise is the combination and levels of expertise of data professionals. Previously, data professionals have been described as requiring a T, I, or Γ-shaped skill sets, encompassing breadth in domain, computation, and methods expertise areas and depth in at least one of these areas (Bloom, 2017; Stanton et al., 2012). In Figure 5.2, the I-shaped data professional has expertise depth in analytics and breadth of expertise in engineering, curation, and domain areas.

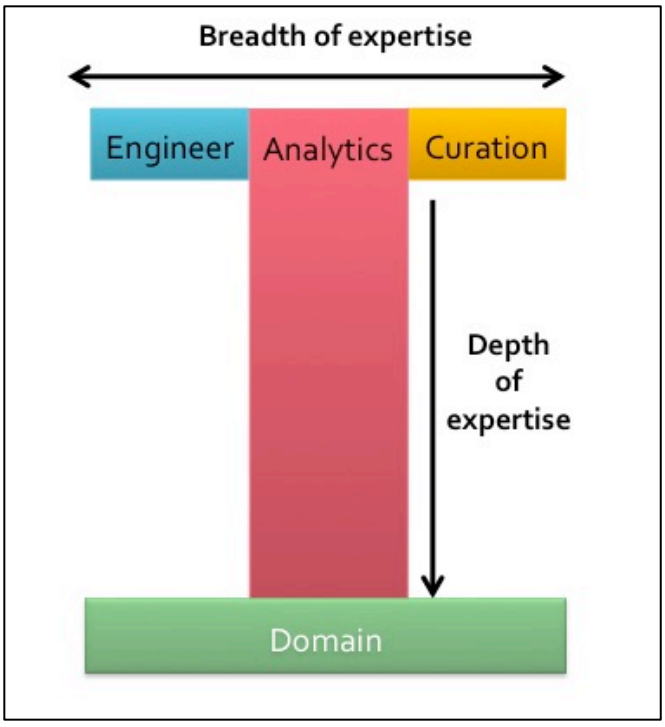

Figure 5.2. I-shaped data professionals based on Stanton, Palmer, Blake, & Allard (2012)

This study found in practice that depth in one expertise category was not enough for data professionals. The participants had expertise depth in multiple areas, resembling the letters of M and N. See Figure 5.3 for a visualization of M- and N-shaped data professionals. The data professional expertise depicted in Figure 5.3 has deep expertise in the areas of analytics, service, and leadership while possessing a breadth of knowledge in the data, research, curation, and engineering.

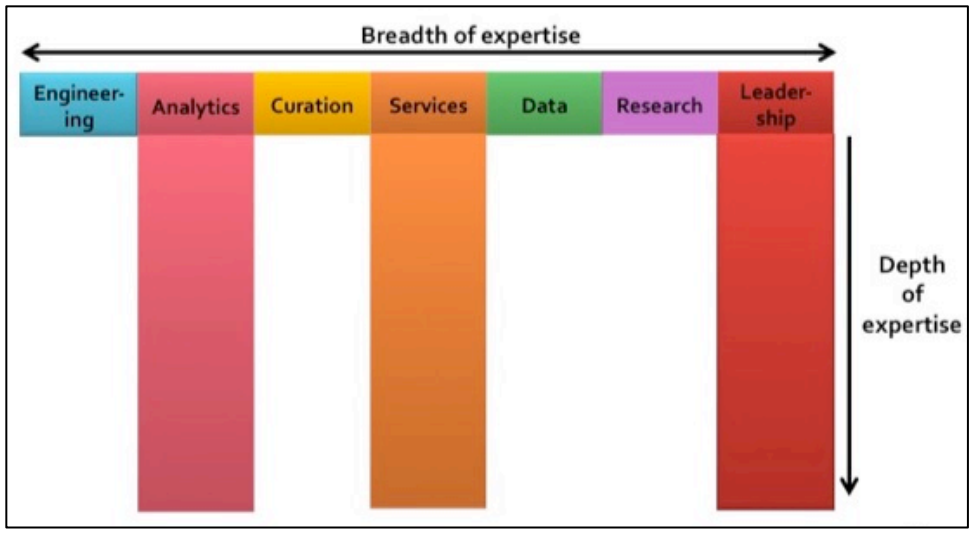

Figure 5.3. Expertise depth and breadth of M-shaped data professional

The NCAR data professionals had multiple data roles in their positions such as Curator, Engineer, or Scientist. The multiplicity of roles required a depth of expertise in relevant practice areas. For instance, the NCAR-Archive Data Engineers served as the Curator, Engineer, and Consultant roles in their positions requiring a depth in the expertise areas of curation, engineering, and service. Similar trends were observed in NCAR-Solar and Climate labs. All the data professionals interviewed exhibited multiple expertise depth. For instance, the Data Scientists required deep expertise in analytics, curation, and services to fulfill their roles of Data Scientist, Curator, and Consultant; while the Data Engineers needed deep expertise in engineering, curation, and services to perform their Data Engineer, Curator, and Consultant roles. With depth in multiple expertise areas matching their roles, data professionals were addressing and solving the data management challenges at NCAR.

The Purdue case told a slightly different story. Purdue data staff often had depth in two or three expertise areas corresponding to their staff roles. For instance, the Data Specialists exhibited depth in curation and service expertise areas meeting the needs of their curator and consultant roles whereas the Senior Data Specialist had depth in curation, service, and leadership

to fulfill the roles of Data Curator, Consultant, and Service Manager. The data services staff possessed depth in two or three expertise areas, allowing them to respond to the demands of data work from their community.

The case differences in combinations and levels of expertise in data professionals reflected the organizational approaches to data staffing and position design. NCAR data professionals were relied upon by scientists and users to provide data support across the research data lifecycle, fulfilling multiple roles and niches for their assigned teams/projects. However, Purdue established a network of data experts across the campus that the library depended on for more depth or additional expertise. The configuration of expertise and roles in data positions is an evolving process at both research sites.

5.3 Professionalization Disconnects and Dilemmas

The story of data professionalization and service development in the research sites drew attention to disconnects and dilemmas (Cutcher-Gershenfeld & Ford, 2005). Previous research has documented a series of disconnects, where there is a gap between what people hope and what actually occurs, and dilemmas, meaning a decision where no choice is superior, each choice has tradeoffs and consequences, but action is required (Cutcher-Gershenfeld & Ford, 2005). Navigating these dilemmas and disconnects influenced the development of data staffing and expertise in the two sites. Through iterative coding of case data, a set of disconnects, dilemmas, and strategies for navigating these harsh realities was identified.

Professionalization disconnects represented the challenges of changing expertise and staffing attitudes and norms in established research institutions. In both research sites, data professionals encountered a similar set of disconnects. Despite the support for data management in the research sites, the career advancement systems have not evolved to include data-specific

tracks. As professionals moved into data positions, they lost an explicit, formally recognized career track in their organization. As discussed previously, the NCAR job classification system recognized scientist and engineer tracks with a ladder of entry-level to senior positions. The missing recognition of data management left data professionals gaining more expertise with no corresponding positions for advancement within the organization, as well as few options for advancement in their career. Purdue data professionals experienced a similar disconnect, where the library did not have data management represented in the job classification system. However, the placement of data professionals into tenure-tracked librarian positions represented a chance for advancement. At the time of data collection, one data professional had received tenure. This disconnect was observed at the supplementary data centers and libraries too. As discussed in section 5.1.2, data-specific tracks did not exist in these organizations, and the advancement opportunities in non-data tracks (e.g., management, engineering, science) shifted professionals away from data curatorial responsibilities.

A second disconnect related to who moved into middle management positions in data services. At both sites, the organizational structure for data services possessed few middle management positions for data professionals. At NCAR, many middle management positions in both science teams and data teams were held by individuals with a research background. The two managers of NCAR-Archive and the senior Data Scientist in HAO held positions with middle management responsibilities and had professional background working in geoscience research. This disconnect was observed in most of the supplemental data centers with the exception of 2 centers where the managers held computer science and engineering degrees but had previous work experience in scientific data management. Similar to NCAR and the data centers, many Purdue data service management positions were held by individuals with deep experience in

research. The Purdue data professionals managing the Purdue-Archive and Purdue-Consult teams had previous experiences as embedded librarians in research projects. The supplemental libraries often employed scientists without a MLS degree or librarians with a science background to lead their data services. While the number of data service manager positions were limited at both centers and libraries, data professionals with research backgrounds, not engineering or traditional library backgrounds, were selected for middle management.

A third disconnect highlighted the loss of traditional professional identities of scientist, engineer, or librarian as data professionals pursued data expertise and work. Throughout NCAR's history, the primary professional types have been the scientist and engineer. At NCAR, the data professional represented a new type of worker, and data services were a new type of work. Data professionals at NCAR often had science or engineering backgrounds and were observed as deviating from their original profession. Conversely, at Purdue, the librarians that moved into data work were able to maintain more of their original professional identity by seeing data services as a new service area for librarianship. Peer librarians recognized their data service colleagues as pursuing a new area of practice similar to recent shifts to digital libraries or knowledge management. While this disconnect was not mentioned directly in the supplemental interviews, both data center and library professionals expressed the challenge of working in a new space like data management where most of their colleagues did not understand their work. This disconnect highlighted the challenges in implementing data innovations and in changing established staffing and professional norms.

Data expertise and staffing development drew attention to the challenges in the whole organization understanding the value of these initiatives. The value of data expertise and work was not distributed evenly across the organization. While NCAR and Purdue data professionals

and teams were excited about new data skills, systems, and tasks, they encountered questions of how data management contributes to science, what data professionals can do, and how their role differed from scientists and other colleagues. This disconnect was expressed in the supplemental data center and library interviews, where data professionals were helping their colleagues understand data work and its importance.

The old adage*, if it ain't broke don't fix it*, relates to the stories of data expertise and staffing development at both sites. At NCAR, staff members were resistant to a few of the data innovations such as the introduction of relational databases, metadata standards, and code libraries. At Purdue, some of the liaison librarians were resistant to new services for research data and their new role as a service access point for data management. This workplace disconnect reflected how the whole organization, not just employees engaged in data services, needed to value and understand data expertise and services. This disconnect was not observed in the supplemental interviews due to the nature of the interview questions.

The organizational disconnects elucidated two underlying dilemmas for data expertise and staffing initiatives at the research sites. A tension between data innovation and organizational priorities was a dilemma at both research sites. NCAR is a science-driven organization with a strong mission to advance atmospheric and climate research. The choice between advancing science goals and future data innovations is a dilemma because there is no clearly superior option. NCAR cannot ignore the immediate science work since that is why NSF has funded them. At the same time, the data services and infrastructure are needed to ensure innovative and reproducible science. A strategy to overcome this dilemma was the NCAR director's support of data management initiatives like DSET. This is an enduring dilemma and will continue unless NCAR can incorporate data into their definition of science.

At Purdue, the data innovations pulled librarians away from traditional services like liaison, metadata, and archiving. Purdue is a traditional academic library with the mission to advance knowledge. This is a dilemma involving a choice to maintain traditional library services or data services. There is no clear choice because the library must continue to provide collections and resources that meet their faculty and student needs. Library data services will ensure they meet the future needs of their constituencies. To address this dilemma, Purdue has disseminated knowledge about data management and services across the library (e.g., through the activities of a Data Education Working Group), established teams dedicated to research data, and integrated research data curation into their strategic plan. These actions have moved data management into the broader organizational priorities and activities and enabled librarians to get comfortable with their new roles in data services. The library has remained relevant and responsive to the campus needs and changing data landscape.

A second dilemma was the tension between centralized and de-centralized governance. As previously noted, NCAR is a fairly de-centralized organization that gives the labs the authority to design teams, services, systems, and practices. However, the NCAR directors are advocating for data innovations and offering resources to support DSET, a team focused on developing organization-wide best practices and data discovery system. The dilemma is that each lab must choose between adopting the innovation from DSET versus continuing to pursue practices and systems that meet their lab needs. This is not an unusual dilemma for NCAR. For instance, a previous attempt at an NCAR-wide data catalog, Community Data Portal, was adopted inconsistently across the labs and even within one science lab. However, DSET set out to solicit lab involvement and feedback through the selection of representatives across all labs and major research areas and to keep the NCAR community aware of their efforts through a

communication plan including progress reports, newsletter updates, and DSET representative report-outs to their labs. Since labs and their members understand the mission and work of DSET plus feel like their science area or data type was represented in the team, the early efforts to adopt data citation practices and engage on metadata activities have been successful. Through leadership support, representation of each lab, and the efforts of DSET members to be open about their work, the team has made significant progress tackling the inconsistency of data services, practices, and systems across NCAR.

At Purdue, a new Dean of Libraries moved data management services up to a library priority. Early on this new priority resulted in mixed responses across the library. A few technology and liaison librarians were excited to develop services in this new area and understood the relevance of data to library work. Initially, several of the liaison librarians did not feel comfortable with their role in data services, as the service access point, and did not want to adopt it. This presented a centralized vs. decentralized dilemma where the Dean was prioritizing data services but the individual units were deciding whether to adopt this new service area. To navigate this dilemma, Purdue started a Data Education Working Group, aimed at librarians developing an interactional data expertise, meaning the ability to converse on data management issues with scientists (Collins & Evans, 2007). Similar to many academic libraries, Purdue encourages professional development, and many librarians have taken advantage of external interest groups, professional association meetings, and workshops to continue to hone their expertise and services. Staff education provided librarians the confidence to fulfill their new data roles and make more informed decisions about data services that they offered. Overcoming this dilemma enabled Purdue to design innovative and responsive data services with the involvement of multiple library units and perspectives.

The similarities of professionalization disconnects and dilemmas across cases highlights general challenges in organizational change for data expertise and staffing development. Data professionalization efforts encountered five disconnects of career advancement systems, promotion to middle management, loss of professional identities, value of data efforts, and staff resistance to change. These disconnects revealed the underlying dilemmas of tension between data innovation and organizational priorities and between decentralized and centralized governance. The sites employed a variety of strategies to overcome these dilemmas including staff education, data team creation, leadership support, open communication about data innovations, and integration of data management into strategic priorities. The professionalization of data work is an evolving process at both research sites. The success of data expertise, staffing, and service developments depends on harnessing the disconnects and dilemmas to support transformational change.

5.4 Summary of Research Data Staffing and Expertise

In this chapter, I compare data staffing and expertise across my two sites. The analysis detailed the two models, in terms of organizational structure for data services, position design and roles, expertise types, and combination of expertise categories. The findings make explicit the prominent elements required to support research data services.

The study results presented two models for supporting research data management. NCAR employed a model placing data professionals in both data and science teams, enabling local expertise customized to each project coupled with deep data preservation expertise available across the organization. In contrast, Purdue built teams of data professionals and took advantage of existing expertise on campus to cultivate a network of additional experts inside and outside the library. This network allowed the library to develop a set of expertise in-house leveraging

additional knowledge and resources in other campus units for research data services. These two approaches to staffing data services emphasized the importance of teams and collaborations and where data expertise resides in the organization.

Over time, data positions and roles at NCAR and Purdue evolved and responded to the data services demands. The data positions comprised a combination of roles – Data Curator, Data Liaison/Consultant, Data Engineer, Data Scientist, and Data Service Manager. The combination of roles in new data positions continued to emerge at both sites. The roles documented the work activities and professional areas of practice for data management and curation. The findings on team structure, positions, and roles can guide research administrators and managers in how to support innovative research data services.

The study findings identified a set of needed expertise types and categories for data professionals. The data expertise types and categories were a set of knowledge, skills, and experiences relevant for data work. The expertise types were established using NCAR interview data and, then, verified and expanded using the Purdue interview transcripts. The categories that emerged represented several areas of practice for data professionals – curation, engineering, services, analytics, leadership, research, and data. These categories are closely related to professional background of data professionals in this study. At NCAR, most of the data professionals had a geoscience and/or engineering background. At Purdue, most data services staff had an LIS background, bringing a deep expertise in curation. These professional backgrounds were present in the emerging data expertise categories and aligned with data roles. As the data profession is still an evolving field, this study contributes the expertise requirements needed for data work blending knowledge from several professions – science, engineering, librarianship, statistics, and management.

The organization context has an impact on the data expertise and roles developed at each research site. NCAR, a fairly de-centralized organization, employed a model of localized expertise where individual labs cultivated their data expertise to meet their scientific data needs. The centralized model at Purdue enabled the library to select types of expertise to develop in-house. These different organizational approaches to expertise resulted in data professionals with breadth and depth in multiple expertise areas. NCAR data professionals exhibited a breadth of knowledge in all the data expertise categories but had deeper knowledge in at least three categories relating to their data roles. At Purdue, expertise depth in 2 or 3 areas were observed in their data services staff. These findings shift the conversation of the skill set of data professionals from one deep expertise area (e.g., T- and I-shaped) to depth in multiple areas of practice. (e.g., M- and N-shaped). My research enhances our understanding of the skill set of data professionals by documenting the multiple areas of depth and expanding the areas of breadth needed (data, research, curation, engineering, service, analytics, and leadership).

The process of data professionalization at each site includes a set of disconnects, dilemmas, and strategies needed to engender organizational change. Data professionalization disconnects revealed the challenges of starting new data innovations in an established organization where tensions appeared in career advancement systems, promotion to middle management, loss of professional identities, staff value of data efforts, and staff resistance to change. These disconnects uncovered organizational dilemmas between data innovation and organizational priorities and between decentralized and centralized governance. Several strategies to overcome these dilemmas were employed across the research sites such as staff education, data team creation, leadership support, open communication about data innovations, and integration of data into strategic priorities. The study results contribute how organizations embarking on data

expertise and staffing development need to value and harness these professionalization disconnects and dilemmas to support innovative and responsive data service models.

The next chapter focuses on strategies and conditions for learning research data expertise.

CHAPTER 6. LEARNING RESEARCH DATA EXPERTISE

NCAR and Purdue like their peer institutions are building research data expertise in order to support and sustain services for data management and curation. Chapter 5 identified two models for data staffing that emphasize the importance of aligning multiple data roles and expertise areas into data positions. Both sites established models where data staff bridged multiple roles and professions, and these staff had cultivated expertise spanning multiple expertise areas. A second objective of this research was to understand how organizations learn to work with research data, focusing on the act of building and sustaining this expertise. Whereas the previous chapter presented results on emerging roles and types of expertise, this chapter documents the learning strategies and influential conditions involved in expertise development. The results are organized into two parts: 1) learning strategies for research data expertise and 2) conditions that contribute to learning. These results on learning can serve as a model for organizations that wish to support professional development.

6.1 Learning Strategies

To offer research data services, organizations need a well-prepared staff knowledgeable in multiple data practice areas. My study identified a set of learning strategies used by both research sites to enhance data expertise. The following sections describe the three overarching learning activities of data expertise acquisition, sharing, and retention. These strategies were identified through iterative coding related to learning in my study data. The three overarching types of strategies – acquisition, sharing, and retention – were informed by typologies of organizational learning processes (Argote, 2011; Huber, 1991). Each institution placed different emphasis on the learning strategies given the organizational context and length of history offering data services. These strategies were executed to achieve different learning objectives

such as managing a new data type or implementing data preservation techniques and best practices. The learning activities were initiated by individual staff members needing to solve a data work challenge and by management addressing service gaps. While emphasizing different areas of expertise development, both research sites made substantial progress in service development using these learning approaches.

6.1.1 Data Expertise Acquisition

Strategies for data expertise acquisition focused on gaining new information about an aspect of data curation or the user community to improve data services or systems. Each research site reported multiple data expertise acquisition processes focused at both individual and group learning. Data staff initiated most of these data expertise acquisition events as needed, responding to new data formats, user needs, data requirements, or technology changes. Table 6.1 is an overview of the data expertise acquisition strategies identified. This section is organized into types of strategies and their outcomes.

Types of strategies

A variety of data expertise acquisition strategies were observed in my interview data. The results reveal a blended set of workplace learning processes: inherited, experiential, grafting, observational, and traditional book/classroom learning. Table 6.1 provides an overview of the expertise acquisition strategies.

Learning activity	Description
Inherited	Learning before organization was formed
Experiential	Learning by doing
Grafting	Hiring new professionals
Observational	Learning from watching others
Book/classroom	Learning by training courses and resources

Table 6.1 Data expertise acquisition strategies

Inherited

As Huber (1991) highlighted in his typology of learning strategies, organizations always begin with knowledge from the founding or that existed before its merger. In both cases, there was a strong, established base of inherited expertise in some areas. At NCAR, the organization started with a deep expertise in atmospheric and climate sciences, their research practices and tools, and working with data types and formats in this domain. Chapter 4 noted the founding of NCAR-Solar lab 20 years before the founding of NCAR, bringing two decades of expertise in collecting, managing, analyzing, and sharing data to the creation of NCAR. At Purdue, the library started with a strong expertise in information organization, preservation, and handling skills before they embarked on data services. Several librarians described learning how to translate library science concepts and theory to research data, which was a new object for them to manage and preserve. Inherited expertise provided a foundation where these organizations and employees were able to utilize an existing expertise and competence while continuing to hone their skills in other areas.

Experiential

Learning by doing or experiential learning (Kolb, 1984) was the most common theme in the interview data. This approach occurred using formal and informal work projects. Several participants at both sites used phrases like “trial and error,” “jump right in”, and “we just did it.” The experience of working with research data created informal opportunities for experiential learning for both individuals and groups. Advances in data formats, simulations, or technologies often initiated this learning strategy. At NCAR, all data professionals reported learning new data formats, programming languages, analysis techniques, or technologies by informal experiments at work. NCAR data professionals emphasized how they learned several programming languages

by picking a common data task and re-creating it in the new language. The Purdue data professionals expressed similar themes of learning data work by doing it. For instance, the Purdue-Archive team experimented with different repository systems as a learning opportunity:

> We stood up D-space, Fedora, and E-prints and said, 'well let's just throw data at these things and learn by doing.'...For our own internal purposes, let's go through the exercise of acquiring, describing, organizing, ingesting, and archiving the data in a repository. We'll do that as a way of learning about the platforms and what are the issues that we're going to run into. (Purdue 101)

Informal work experiments and assignments were a valuable learning tool for data professionals, allowing the deep exploration of data problems and a platform for trying new approaches or tools.

Formal work assignments represented another mechanism for experiential learning. Purdue placed librarians in research projects as a means to learn about data management and curation. These opportunities allowed librarians to work with scientists and to gain first-hand experience in handling data and the challenges in curating data sets. As one member of the Purdue-Consult team described her/his experience:

> We [librarians] started working on [science] projects that were…a pathway to build relationships, sort of understand deeply what people were doing and, then, if something came up the next time, you could maybe repurpose what you had learned or what you had done. (Purdue 110)

While the library placed librarians into science projects as a valuable learning tool, the librarians also learned from library grants investigating data curation service issues. The library has a history of successful research projects such as Data Curation Profile, Data Literacy Curriculum,

and DataBib (Brandt, 2013; Witt, 2012). These projects provided avenues for staff to investigate their questions about data services and learn from doing research as illustrated by the quotes below:

> We realized that we could interview these people to really sort of figure out if there was data common across different disciplines…[the] first data curation profile grant where we went out and interviewed... and I mean we had lots of conversations and it went to lots of different places so that's when I think we realized we need to be doing more of this. (Purdue 110)
>
> ... so part of the reason we're doing the study of organizational models for libraries providing data services is because we here at Purdue don't have it figured out yet. (Purdue 101)

Purdue has an annual retreat designed to promote reflection within the data teams from these work projects and continue the learning. Assignments stemming from research projects provided a formal means for data professionals to learn about data practices from science projects and data services from the library grants.

A similar theme of formal projects was present in the NCAR-Archive team as a means of staff development and exploration of new service features. These projects responded to emerging data needs recognized by the team of data professionals. Incidents of new learning associated with special projects included data transfer technologies, website usability, and metadata standards. For instance, the manager described a metadata project:

> We've got all these different data sets, and they have quasi-standard metadata but not quite. So I went to [employee name redacted] and I said, '[employee name redacted] could you help us work things up and get all the metadata standardized

> across all of the data sets' and he said 'yah, I think I'd like to do that.' He jumped in and that [metadata service] has just evolved into what it is now. (NCAR 211)

This project improved the metadata quality and compliance with standards as well as enhanced expertise in geoscience metadata standards in this team. These special work projects enabled one or two team members to experiment with a new standard, service, or technology and learn how it relates to research data and integrates into local data practices. This team also built an additional learning activity into these special projects by asking data professionals to reflect on lessons learned in their annual performance review. Formal work projects coupled with reflection were successful learning strategies for data professionals at the two research sites.

Grafting

A third mechanism for gaining new expertise was grafting or hiring new professionals with necessary skills sets (Huber, 1991). Both research sites reported this as a way to bring needed expertise into the team and organization. Data service managers elaborated on the importance of hiring new professionals to fill a missing or deepen an expertise area. For instance, Purdue recognized a missing competence in handling big data among their staff, so the library hired a data professional with a background in large-scale data and computation. The NCAR-Archive is another example where they hired a data professional with expertise in working with oceanographic data to complement the other staff with expertise in meteorology data and simulations. As the data professional noted their expertise: "I came as an expert oceanographer...They [NCAR-Archive] had a group of I think 7 software engineers at that time. They saw the need of having an oceanographer involved" (NCAR 211). The hiring of new data professionals provided opportunities to acquire specific skill sets that complemented internal, existing expertise. New data professionals can also bring expertise in unanticipated but relevant

areas. For example, NCAR-Archive hired a second data professional to expand the oceanography expertise but this individual added database expertise into the team: "I'm the one who introduced database processing into the group, so I basically created, of course with other people's help, the RDA MS- its RDA [Research Data Archive] Management System" (NCAR 201). This data professional arrived at a time when database structures were advancing and becoming more prominent in science. New staff members brought expertise in emerging or missing practice areas, offering a useful technique for expanding competence in a team or organization.

Observational

A fourth data expertise acquisition process was observational learning (Bandura, 1986), comprised of learning from colleagues and professionals in other fields. This learning strategy was effective in both research sites. At both sites, data professionals were able to turn to co-workers and professionals across the organization to learn about standards, practices, new technologies, or domain knowledge. NCAR-Climate professionals drew attention to learning about more domain knowledge from local scientists and data management from NCAR-Archive or library professionals. In NCAR-Solar, data professionals also reported learning about data citations and identifiers from the NCAR Library's Research Data Scientist. NCAR data professionals were a valuable resource for each other. At Purdue, learning from colleagues was noted in two directions: librarians learning from data professionals and data professionals learning from librarians and service partners in other departments. The data professionals in the library's data teams were a resource for liaison librarians and other library professionals to learn more about data management and preservation for their designated community. The data professionals reported learning from their librarian colleagues about disciplinary or metadata practices, compliance from research office professionals, and domain-specific knowledge from

campus researchers. Co-workers were a valuable resource for data professionals to learn more about domain or data management best practices, standards, norms, and technologies.

Observational learning occurred by looking outside of the organization for inspirations and insights into data work. Most data professionals elaborated on how observing data practices, and systems in other disciplines or data centers was a useful learning technique. These observations provided new insights into data issues such as data movement, discovery, or metadata. All NCAR data professionals emphasized learning about data challenges or solutions from personnel in other geoscience data centers. Looking to professions outside of the geosciences, NCAR data professionals also have gained insights that related to their data work. For instance, NCAR-Archive professionals were inspired by the notion of movie credits as a model for providing acknowledgements of multiple contributors to a climate simulation. By analogy, the credit of several contributors and roles informed an attribution and acknowledgement framework for climate model citations (Hou, Betancourt, & Mayernik, 2015). Similar trends were observed at Purdue. Purdue data professionals emphasized the importance of observing researchers in the field and learning from peers at data-focused meetings or conferences (e.g., Research Data Alliance, Earth Science Information Partners) to learn more about data management practices and solutions. These new insights have informed the library's data practices and repository service. The study findings documented observational learning as a strategy for data professionals to learn about research data expertise.

Books/Classroom

A final theme of data expertise acquisition was learning from traditional book and classroom approaches. At both research sites, data professionals reported attending formal courses both as part of degree programs and as part of continuing education. Data professionals

have the option to attend local workshops, trainings, and seminars to advance their expertise too. Local offerings focused on data-related topics and domain topics. Formal staff trainings were present in the NCAR-Solar and Purdue-Lib groups. The NCAR-Solar lab offered staff training for new data professionals focused on data quality and processing skills. Purdue's e-Data Task Force created a libguide for liaisons and other librarians to learn more about research data and data management. In addition to trainings, books and online resources served as useful learning tools. At NCAR, data professionals were often expanding their skill set in terms of programming languages or tools using online books or resources such as *W3Schools* and *Stack Overflow*. Purdue data professionals kept abreast of current literature on data curation. Educational programs and resources helped data professionals to expand and update their skill sets.

Outcomes

At NCAR, an emphasis on data expertise acquisition led data professionals to pursue new knowledge and experiences working with research data to address the challenges of data management, archiving, and sharing. The mix of strategies made it possible for data professionals to update and expand their skill sets to meet the needs of their organization and user communities. Data professionals cultivated a deep expertise in data, research, analytics, and engineering areas specific to their data work through various learning strategies and strong relationships with colleagues in their organizations and communities. Most NCAR data professionals had a geoscience and/or engineering background providing a foundation to expand their expertise in other practice areas such as curation or service. These strategies worked especially well for a domain research institution with a strong connection to the user community and a strong commitment to continuous improvement.

Purdue data professionals focused learning strategies on extending the traditional

librarian skill set to working with a new object, research data. Data professionals learned to apply LIS concepts such as collections and description to data sets and expanded their expertise to data, research, and services areas of practice. Rather than cultivating all aspects of research data expertise, the library leveraged existing deep expertise in engineering, analytics, and research compliance from other campus units. This mix of strategies worked well for a library relying on a network of strong collaborations across campus.

6.1.2 Data Expertise Sharing

As organizations gain expertise about data management, learning efforts can focus on data expertise sharing to distribute the expertise into other units and/or across the organization. Two strategies of sharing appeared in the interview data from both research sites: task forces and peer-to-peer learning. These learning activities focused on transferring expertise of concepts, requirements, standardized practices, or technologies related to research data work. The strategies were initiated from various levels such as from workers, teams, and management. The intended audience for this learning was primarily data professionals and scientists but extended to other professionals in the organization. This section reviews the two learning strategies and their outcomes.

Types of strategies

Task forces

Local task forces were a popular transfer strategy that often was initiated by data professionals or other employees. Employees formed these cross-unit collectives to provide a forum for sharing lessons learned, emerging data standards or trends, technologies, and best practices, as well as designing new software. These groups were time-limited and driven by a mission to educate other staff and/or produce a shared approach or system for a specific data

challenge. A key feature of these groups was cross-unit membership, bringing together professionals with different expertise.

NCAR had a series of data-focused task forces and groups throughout its history (see Chapter 4) that enabled expertise sharing. These groups focused their efforts on data citation, shared data discovery system, and organization-wide best practices. While the mission was to address a local data challenge across all the labs and teams, these meetings offered opportunities for NCAR data professionals to interact and learn from each other. Often, data professionals working in science teams were not aware of data professionals and their activities in other labs. The local groups spurred seminars, demonstrations, or shared tasks where data professionals could share and learn from each other. For instance, the current Data Stewardship and Engineering Team (DSET) allowed data professionals from different labs to interact and fostered opportunities for staff to learn from each other. The NCAR-Solar lab DSET representative described the meetings as allowing her/him to seek advice from seasoned data professionals on best practices:

> I asked this question the other day [at the DSET meeting] and was like 'does anyone else serve their data through Drupal' and they were like 'No, we have our own web servers in our labs that is how we serve our data.' (NCAR 101)

The second NCAR-Solar DSET representative was excited to serve on this committee as a way to learn more about scientific data management since s/he was new to the field and hoped the experience would be similar to how s/he utilized her/his role on the local Web Advisory Board to learn more about web design and usability. Task forces were effective for social learning and for cross-unit progress on shared definitions of terms like data (Baker, Mayernik, Thompson,

Nienhouse, Williams, & Worley, 2015), on community data portal, and on a common set of data citation practices (Mayernik et al., 2012).

Similar task forces formed at Purdue to explore and learn about data management and curation service needs. The early work of the library-wide e-Data Task Force, started by library data professionals, led to the successful campus support and funding for the PURR data repository development, resulting in the motivation for the campus-wide PURR Working Group formation to guide the development work. Both these groups provided opportunities for data professionals and librarians to learn about data repository service through social experiments with repository technologies and by sharing their expertise. A third task force (Data Education Working Group) was initiated by liaison librarians that wanted to learn more about their new duties in research data services, inspiring a series of seminars and libguides of resources to share expertise about data management. Furthermore, a Senior Data Specialist illuminated the importance of the social learning dimension of participating in local groups on research data: "It was like learning about the business of research, learning how to interact with people about research, and a lot of that was about data [management]" (Purdue 110). The Purdue groups fostered sharing of expertise and concrete outcomes such as a shared repository system and practices, library policies for data, and new data staff roles, as well as cultivation of an internal data community of practice in the library.

Peer-to-peer learning

A second sharing technique was peer-to-peer learning, where data professionals shared their expertise with co-workers and colleagues in their team and other units. This activity comprised formal and informal approaches in that it was not always planned or intentional. Hallway or lunchroom conversations sometimes resulted in data expertise sharing between data

professionals and/or scientists.

At NCAR-Archive, the Data Service Manager built the practice of peer-to-peer learning into the annual performance appraisals. These data professionals are expected to share their expertise with their co-workers. As the Data Service Manager described the performance appraisal process: "I always stress the fact that they have to, that they are responsible for educating, supporting, and training the other staff on how to do certain things…For example [Name redacted] teaches people how to do many different things with the database" (NCAR 211). While this was a formal duty in NCAR-Archive, the learning responsibility was observed in other labs as well. Several data professionals reported learning from other professionals in their labs on such topics as data practices, web delivery, or metadata, as well as from the NCAR Library's Research Data Scientist on data citation and identifiers. An NCAR-Climate data professional noted learning from NCAR climate scientists in her/his lab and from a data professional in the NCAR library:

> I get to eat lunch with some of the best scientists that are in this field…sometimes when they get into a big science discussion I just sit there and listen and sometimes it's over my head but you learn that way…and then I've talked with people like [Library's Research Data Scientist name redacted] and others who are much more familiar with the library science and data science aspects, so it's been a lot of learning. (NCAR 206)

While NCAR is divided into three campuses, each campus has a variety of professionals and has a shared cafeteria where staff from different projects or units at the campus can interact. Furthermore, I. M. Pei, a renowned architect, designed the physical layout of the Mesa Lab to promote employee interactions.

Peer-to-peer learning occurred primarily among the Purdue data professionals but also

across the library staff. Each library data professional brought a particular background and expertise into the library. This expertise was shared with data professionals, liaison librarians, and other librarians through shared work assignments and informal conversations, among other opportunities. Most library interviewees commented on learning from their peers in the library. As the Senior Data Specialist commented on peer-to-peer learning:

> Our charge is to learn about data management…Each of those [data professionals] has a specialization so [name redacted] is big data, [name redacted] is metadata, and [name redacted], her/his background is Anthropology…we have the data expertise that we're trying to build and share, the specialization expertise that we are trying to build and share. We share that with the liaisons so we [data professionals] are sort of central but we want to move outward [into the library]. (Purdue 110)

The movement of data professionals into data teams and into one physical, prominent library location allowed more opportunities for the data professionals to interact and share their expertise. The prominent, new location in the library improved the visibility of data professionals on campus, making them easier to find and to foster peer-to-peer learning chances.

Outcomes

Data expertise sharing activities were motivated by a desire to develop NCAR-wide knowledge, definitions, practices, and systems, and an individual goal to continue learning and improving data management skill sets. Many data professionals expanded their expertise from their initial background in geosciences or software engineering through peer-to-peer learning opportunities to other areas. The participation in task forces also enabled data professionals to share best practices and lessons learned, cultivating a shared expertise for data management. The expertise transfer strategies worked well for NCAR, an organization with a combination of

seasoned and nascent data professionals, where sharing expertise happens in both directions.

Purdue's desire to distribute data management expertise across the library led data professionals to share their expertise in informal conversations and formal task forces. Local task forces resulted in common understandings of data management across the library and even prepared liaison librarians for their role as the first point of access to research data services. The diverse backgrounds of data professionals provided opportunities for expertise sharing among staff members.

These data expertise sharing approaches were effective for organizations like NCAR and Purdue that had existing expertise and a desire for more shared approaches to data management and curation. The common desire to offer high quality data services motivated data professionals to continue sharing their lessons learned and to seek new expertise.

6.1.3 Data Expertise Retention

Once a specific expertise has been cultivated, organizations may direct their efforts to retaining this information and preventing loss. While data expertise retention efforts were present in both research sites, NCAR had implemented more of these strategies than Purdue. These differences may be the result of a longer history in offering data services and in building a well-trained staff at NCAR. These efforts emphasized the retention of data management and curation expertise and the development of shared resources containing this expertise such as data resources and practices. The section describes the strategies, case results, and their outcomes.

Types of strategies

Documentation

Data expertise retention emphasized the development and preservation of data management documentation such as technical reports and guides. These resources captured

expertise for current and future employees to use as learning resources. At NCAR, the DSET and DCite task forces deliberately produced technical reports that are preserved in the NCAR digital repository due to the lack of documentation on lessons learned from previous data groups. These reports are accessible to NCAR employees and the general public providing best practices, new insights, and recommendations learned by these task forces. At Purdue, the Data Education Working Group designed and maintains a libguide, a web resource, for educating librarians on data management and containing useful resources such as a glossary, bibliography, survey of tools and of funding requirements, sample data management plans and data curation profiles, and data-related publications from Purdue. The Purdue data professionals published several reports and articles documenting their service design, decisions, and lessons learned in the scholarly literature. These valuable resources capture the expertise of data professionals as well as serve as a learning tool for other employees and broader audiences.

Practices

The development of shared practices was a strategy for data expertise retention at both sites. Shared practices were available for all employees to utilize, and many new data professionals benefited from a systematic approach to data processing or archiving, where the learned data-related standards for metadata or archiving were embedded into the practice. For instance, the NCAR-wide DCite group produced a set of recommended practices for digital object identifiers (DOIs) and data citation documented in the group's technical report. The DCite practice recommendations were adopted across NCAR instilling the expertise of citations and identifiers into research teams and their workflows. At Purdue, shared repository practices were a retention mechanism for preservation expertise. From the Archivist, the Purdue-Archive staff learned about preservations techniques and standards and designed data repository practices and

workflows steeped in preservation expertise and standards (e.g., OAIS, ISO 16363). Shared practices allowed the expertise of data professionals to be retained.

Outcomes

Based on my observations of the two sites, the implementation of data expertise retention strategies signified maturity in data service development and organizational learning. Data expertise retention efforts benefited current and future employees that may learn from these resources steeped in data expertise and protected the sites from knowledge loss due to staff turnover or retirements. These approaches worked for these research institutions because of the value placed on knowledge discovery and professional development.

6.1.4 Summary of Learning Strategies

Both sites established a program for learning research data expertise including a variety of learning strategies to acquire, share, and retain expertise. The data expertise acquisition activities focused on enhancing the current skill set of data professionals to meet the demands of their data roles as well as enabled other professionals to learn about data management and how it might relate to their work. The blend of inherited, experiential, grafting, and observational processes with traditional book and classroom learning worked well. Since these two sites had acquired data expertise and competence, the learning strategies included approaches to data expertise sharing and retention. Local working groups and peer-to-peer learning represented formal and informal approaches to data expertise sharing to distribute the expertise across the organization. Finally, the embedding of expertise into shared documents and practices enabled the organization to retain data expertise and thwart knowledge loss. The effectiveness to share and retain expertise has been observed to impact organizational learning in the literature (Argote, 2013). Both research sites established a set of learning strategies to build a well-trained staff

capable of supporting data services.

As previously mentioned, the supplemental data center and library interviews did not include questions about learning strategies; however, participants inadvertently reported some of these learning processes. The most prominent trend for expertise acquisitions was experiential learning in libraries (5) and data centers (10). Less frequently mentioned strategies in data centers and libraries were book/classroom and grafting for expertise acquisition and peer-to-peer learning for expertise sharing. Data center and library participants noted the challenges in utilizing grafting strategy because applicant pools lacked qualified candidates with a mixture of domain, curation, and technical expertise areas.

6.2 Conditions Impacting Learning and Expertise

As organizations are embarking on learning, the process of organizational change can be inhibited or enabled by organizational or environmental factors (Argote, 2013). For the research sites, a set of conditions emerged in the history of data expertise and staff development. A number of these factors have been well documented in the literature: organizational culture (Edmonson, 1999), resource allocation (Kimberly & Evanisko, 1981), leadership support (Yin, 1977), alignment with stakeholders and their interests (Wagner, 2007), among others. In this study, three prominent conditions for learning emerged from the cross-case analysis: 1) sphere of influence, 2) local data community of practice, and 3) visibility of data work.

6.2.1 Spheres of Influence

Influence and agency affect the ability of employees to spur service innovations, defying established institutional norms and practices (Garud et al., 2002; Maguire et al., 2004; Raven, Schwarzwald, & Koslowsky, 1998). This study observed that data professionals with extensive knowledge influenced the development of data expertise and services. When data workers were

able to garner support to change norms, this resulted in new staffing, expertise, or services with varying degrees of impact across the institution. The summary of case results revealed data professionals influencing three spheres – 1) organization, 2) unit, and 3) project - impacting the story of data staffing and service development.

The first type of agent was an employee with an organizational sphere of influence, enabling him/her to initiate changes across the organization. These actors were able to realize opportunities, motivate employees and stakeholders, and allocate resources to these data efforts (e.g., staff, computer systems). Often, the employee had a deep expertise in scientific data management issues and/or held an administrative position that granted them agency and authority across the organization. The NCAR employees classified as having an organizational sphere of influence in this study were the NCAR director, NCAR assistant director, and NCAR Library's Research Data Scientist. These professionals advocated for data management, enabling key outcomes such as the DSET formation with allocated resources, adoption of NCAR-wide data citation and identifier practices, and support for a local data community of practice. While the NCAR executive positions have the formal authority to influence the organization, the Research Data Scientist was an interesting agent that due to her/his deep expertise in data curation garnered the respect and support of peers to innovate NCAR-wide data services and expertise. At Purdue, the Dean of the Library was the employee with the ability to initiate university-wide changes like the creation of research data services. The Dean had a deep knowledge of scholarly communication and academic research issues. This position granted her/him access to campus administrators, a campus view to identify the need, respect to garner their support, and the authority to change library priorities, motivate librarians and campus partners, and allocate resources. All four examples became advocates for research data

management early in the tenure of their positions and were effective in implementing institutional changes in practices, attitudes, or norms.

Data professionals with a unit-level influence were effective in producing changes in the lab or library, impacting many data workers and research projects. These employees often held middle management positions, offering them authority over a unit's priorities and resources as well as the ability to motivate unit employees. At NCAR, three examples included the two employees that held the position of NCAR-Archive Data Service Manager and one Data Scientist in NCAR-Solar lab. These employees had long careers in scientific data management and had previous research experience, enabling them to build "contributing expertise," that is, the ability to contribute to research projects in substantial ways (Collins & Evans, 2007). Using their extensive research data expertise, these workers were able to motivate their units to adopt changes such as staff trainings and data service enhancements. At Purdue, four data professionals emerged as exhibiting a library level of influence: Senior Data Specialist, Repository Director, and two Liaison Librarians. All four examples were able to understand the changes needed for data services, gain the respect of their peer librarians, and garner support for the service innovations (e.g., data repository, data roles for liaisons). These agents were involved in the early data efforts at Purdue and had been embedded previously in research teams. These experiences provided a deep understanding of data management and curation. The two Liaison Librarians were a surprising agents to be influencing library-wide changes given their normal jurisdiction is only for services to a department or discipline; their peer liaisons recognized them as experts on data issues due to their research experiences. All seven examples had extensive knowledge and experience in data management, granting them influence over their peers and some agency from their superiors.

The final type of data professional had influence over a project or service, exhibiting the ability to make decisions for data work, learning, and services related to their work assignments. These data professionals were placed in science teams or were assigned to a specific data service, working in the field or in the lab. In contrast to the data professionals with organization and unit spheres, these workers did not have the ability to allocate resources and their sphere of influence was on the performance of a single science project or service. At NCAR, four data professionals in the NCAR-Climate and NCAR-Solar labs impacted their team norms. These data professionals were supervised by scientists or software engineers who were often unfamiliar with data management and what the work entailed. As a NCAR-Climate Data Engineer describes the situation: "One good side about kind of falling in between the cracks and having the experience that I do is having fairly wide latitude of how things get done, which is nice. Because when I talk about various aspects of data management and data science I'm the only one…that has that knowledge. They just say, 'Whatever [interviewee name redacted] wants, that's fine'" (NCAR 206). Data professionals with a project-level influence and specialized expertise were able to recognize the needed changes, convince their supervisors, and implement these changes. For example, they accomplished changes to data practices (e.g., metadata for climate models), service enhancements (e.g., training for scientists), and their own professional development (e.g., metadata, provenance, analysis techniques). At Purdue, workers with a project sphere of influence included Liaison Librarians, Data Specialists, and Data Curators. They were on the direct front lines of data services, giving them first-hand knowledge of the needs and issues. While their positions often had limited agency, they were still able to engage and motivate their peers and supervisors to implement changes within their own service team. A few examples of their influence included the formation of the Data Education Working Group by liaisons and

implementation of new trainings on big data or literacy.

Data staffing, expertise, and services need to grow and change to keep pace with data management trends. In an organization, professionals engaged in data services are working at multiple levels from management to the frontlines, exhibiting different degrees of agency and spheres of influence. At both cases, a few data professionals emerged that were able to express more agency than typical in their position due to their deep expertise. This extensive expertise enabled these professionals to encourage their colleagues to adopt data services changes at organizational, unit, and project levels. Data professionals developed this extensive expertise as legitimate peripheral participants learn in communities of practice (Lave & Wenger, 1991), where they learn by performing the work. In this study, the data professionals did not have a formal degree in data management or science but become more familiar with data techniques and concepts through doing the data tasks. By participating in research projects, the Purdue data professionals were able to gain insights into research process, data practices, and issues in data work. Due to these experiences, they developed the language for data management, enabling them to communicate with researchers and to gain respect and recognition as data experts from their peers.

When data professionals did not have the ability to modify their services or practices, data innovations were inhibited. An NCAR-Solar Data Scientist compares a data professional without the ability to change services or explore learning in new areas as “a lame duck congress when you can’t get anything done” (NCAR 101). Misalignment between agency, influence, and data positions spurred inconsistencies in the data services offered across the organization. Multiple NCAR data professionals working in science teams described the inequities in research data services in their labs. A few groups had mature data management and archiving services like

NCAR-Archive with data professionals impacting service and unit innovations, while a few teams had data professionals maintaining inefficient, out-of-date data systems or no professional focused on data management. A similar theme was documented in data professionals and services at Purdue. The Purdue-Consult and Purdue-Archive teams had members innovating data practices and norms at service and unit levels, resulting in responsive, high quality data management services, while the services from Liaison Librarians varied across disciplines, where there were no repercussions if a liaison did not provide data consultations. The sphere of influence of data professionals is an important condition for data services, learning, and professional development.

6.2.2 Local Data Community of Practice

Both research sites cultivated local research data communities of practice. The community of practice (CoP) theory of social learning explains how groups form with shared practices and expertise learned over time as group members pursue their shared interest (Lave & Wenger, 1991; Wenger, 1998). This theory emphasizes learning as both individualistic, where humans have agency to act and make decisions, as well as collective, where individuals learn by community engagement and practice. CoP has three elements: shared interest, community, and practices. Both research sites established local research data communities of practice.

The process of learning about and enacting research data expertise has evolved slowly to form a CoP for data management at NCAR. The formation of a data-focused, organization-wide team and task forces, and the establishment of a Research Data Scientist in the NCAR Library, have been pivotal in the development of a local CoP. In the early days of NCAR, science teams and scientists had to manage data, but data work was considered a means to the end. The formation of NCAR-Archive, a data archiving team, brought together a group of professionals

with the shared interests in quality data curation and access. This group worked on shared activities of data processing, software development, and dissemination, allowing a community to form that shared expertise, lessons learned, and experiences and that worked together to learn about best approaches or software, to resolve problems, and to create shared data practices, systems, and expertise. The Data Service Manager encouraged individual and collective learning pursuits. While the NCAR-Archive team allowed social interaction and learning to form the initial CoP for research data within the organization, the formation of NCAR-wide task forces on research data expanded this CoP to include data professionals in science teams. For many data professionals, these task force meetings were the only opportunities to interact with other NCAR data professionals. These task forces enabled professionals with a shared interest in data management to meet and to work on activities to develop shared practices and systems. Another anchor of the CoP was the hiring of the NCAR Library's Research Data Scientist. This position was located in a UCAR unit serving all the science labs, giving this individual an opportunity to identify common data problems and connect professionals interested in these challenges across the labs to share best practices or recommendations among those facing similar challenges. In addition to facilitating these inter-lab connections, this professional co-founded two NCAR-wide data-related task forces (DCite and DSET) and engaged data professionals in research grants related to data curation. This individual played a valuable role in expanding and strengthening the NCAR research data CoP. In particular, the NCAR CoP has cultivated shared data identifier and citation practices, developed an online data catalog (e.g., DSET Search and Discovery System), produced technical reports, and identified data experts that serve as resources for new data professionals.

Similar to NCAR, Purdue has cultivated a local CoP, involving primarily librarians, with a few members from elsewhere on campus interested in data management and curation issues. This community emerged from data-related task forces, as previously described. These task forces formed to address a data question or problem, bringing together professionals with a shared interest from different library units. These groups provided a common work assignment for data professionals to interact, learn, and share expertise. These groups produced a set of shared learning resources, including technical reports, trainings, and libguide. This CoP continues to grow as Purdue extends research data responsibilities (e.g., liaison roles) across the library.

A local data-focused community of practice offered data professionals in different teams and units a network of colleagues with shared interests as well as opportunities for social engagement and learning. The communities fostered the development of a common worldview on research data, practices, and systems. As data professionals contributed to the local data practices and expertise, the community continued to update these practices and knowledge as well as to serve as a valuable resource for new data professionals. These results highlight the positive impact of communities of practice formation. A data-focused CoP is an important condition to enable research data expertise development.

6.2.3 Visibility of Data Work

Recent attention to data management and sharing has highlighted an ignored part of the scientific process related to working with research data and elevated research data to a key part of the modern scholarly record. Even so, research site interviews show that the invisibility of data work is changing slowly. Invisible work refers to "the expertise often hidden from view," emphasizing the visibility of certain expertise, workers, and tasks in an organization (Star &

Strauss, 1999, p. 11). For example, Star and Strauss (1999) describe the invisibility of nurses and care work, often overshadowed by the physicians and medical interventions. Data professionals have been visible members of research teams; however, the curatorial work with research data itself has been invisible. This section explores the visibility of data work as a factor in the development of data services and staffing at both research sites.

Multiple NCAR data professionals drew attention to the invisibility of their work and the changes in visibility over time. In the descriptions of the science projects, I often heard two categories of workers – scientists and others. The other grouping included an array of non-scientist professionals and skill sets such as data management, software engineers, and instrument operators. An example is a participant's description of a climate modeling project staff as "…the scientists here and their collaborators which is like everybody else" (NCAR 208). The combination of all non-scientists into one category makes the diversity of work and roles invisible. Following on this theme, several data professionals noted the lack of recognition for data work and what data professionals contribute to science. The NCAR-Archive Data Service Manager noted how easy access to research data hid the complexity of data preservation:

> In some ways, we're victims of our own success. If the RDA [data portal] just sits there, a person comes and in an instance they find what they need. Then, they're [users are] gone off doing their research. They [users] probably don't even think that there are a lot of people behind that. It's like when you go to buy a book at Amazon and just two clicks. (NCAR 211)

Moreover, an NCAR-Climate data professional reported that a mistake makes data work visible to climate scientists: "If I don't do it or I do it wrong, we know about it right away. Somebody says, 'this data set is all messed up.' There has definitely been a greater appreciation of the work

that I do" (NCAR 206). Data professionals emphasized that observational scientists had more recognition of data work and its challenges given their hands-on experiences working with data in comparison to scientists that rely on existing, publicly available data sets like modelers, technologists, and demographers.

The absence of data roles and professionals in the scientific product contributors and organizational systems was evident at NCAR. For example, data professionals are often not acknowledged in the conference papers, technical reports, and publications disseminated from the research project to which they contribute, masking the role they play in atmospheric science. A few scientists-turned-data-professionals worried about their short vitae and future career prospects given the practices toward credit. Furthermore, the NCAR job classification system lacked formal recognition of data management. The job categories and career ladders for data management were absent. Many data professionals were hired as software engineers, and this shoehorning of new data responsibilities into an existing category resulted in a mismatch of skill levels and position levels. As an NCAR-Climate Data Engineer noted: "Some of the stuff I do would still be considered like a very low-level sophomore engineer like running scripts and running jobs on these machines. But then I get asked to these conferences for the NCAR view on data and participate in things like DCERC and that's like a high-level managerial. So there is this kind of mismatch between some of what I do and some of what I'm asked to do" (NCAR 206). This engineer further expanded on how NCAR does not understand data work: "And NCAR is in some ways still trying to figure out what kind of role I fill in the organization because all this data stuff at least in terms of my work is still kind of in a funny gray area" (NCAR 206). Over the history of NCAR, the visibility of data work has changed. A NCAR seasoned data professional commented on the increased respect for data management and curation: "I do think

there is more respect for the data now. I mean many fields use data, digital data, digital collections. So there is more respect for the data and what it brings to the table in the scientific field" (NCAR 211). While the evidence suggests that the work of data professionals is becoming more visible to the scientists, the visibility continues to be a challenge in the scholarly record and organizational systems.

The visibility of the Purdue library as a partner in research changed over time, shedding light on the important expertise librarians bring to data management. The early exploration of data services was met with questions from scientists and campus administrators on what role the library could play in terms of research data. This quote from a library administrator illustrates the visibility of science and engineering and invisibility of the library:

> "At a University like Purdue, science and engineering are the 900-pound gorillas. We [the library] needed to really align ourselves with them in order to show our relevance and our importance. The provost and president tend to be engineers or scientists, so to get the support from them we needed to be able to show that we were not just passive, but that we were active participants [in research]." (Purdue 102)

As funding agencies were requiring data sharing, the Senior Data Specialist described how scientists were seeking help with data management and asking the librarians: 'Can you do something with [our] data?' (Purdue 110). While researchers were turning to the library for help, they did not initially recognize the library's curatorial expertise and how it applied to research data. The increased visibility of librarians in the data management arena enabled the ability to gain support for the creation and adoption of data management and literacy trainings, resources, and repository services. In turn, librarians were invited to serve on research projects as data

consultants, offering opportunities to learn more about data curation needs and challenges. Purdue data professionals have published extensively on their data service model, lessons learned, and decisions and are recognized leaders in library data services.

Examining the visibility of data work sheds light on how the attention to data activities and expertise impacted the data innovations. At NCAR, data work is moving from invisible to visible, as data becomes a prominent product in the expanding, modern scholarly record. While data professionals reported increasing respect for data management and its contribution to science, the organizational systems for job classification and practices for credit do not reflect the efforts of data professionals. At Purdue, the library has been gaining recognition as a campus partner in research data management, emphasizing the role of the library and value of the curatorial expertise of librarianship. The increasing visibility of data work fostered the legitimacy of data curation expertise, respect for data professionals, and collegial support for data staffing and service initiatives.

6.2.4 Summary of Conditions

In the process of developing research data expertise, organizational learning and change was enhanced by a set of conditions at the research sites. Three prominent conditions of sphere of influence, local data community of practice, and visibility of data work provided the impetus for professional development of data workers. Data professionals exhibited multiple levels of influence within an organization. These spheres of influence enabled data professionals to modify practices, expertise, and staff roles at an organization, unit, and project levels.. Data professionals that expressed a larger influence than their positions allowed relied on their extensive expertise and respect of their colleagues to produce these data changes. The cultivation of a local data-focused community of practice allowed data professionals to meet other

employees with a shared interest and provided a platform for data expertise sharing and developing shared data practices and systems. Data professionals learned as legitimate peripheral participants, where the performance of data tasks cultivated a deep knowledge. A final factor was the visibility of data work, emphasizing the specific expertise and activities of data professionals and their contribution to science. These three conditions resulted in emphasizing the value of data activities and continual professional development of data staff.

I noted previously that the supplemental interviews did not ask questions about learning conditions, and therefore I am unable to confirm these NCAR and Purdue findings on conditions. A few librarians discussed the importance of building a community for research data management in the library and across campus to enhance expertise sharing. Some of the data center and library professionals had found data communities of practice in professional and international groups like Earth Science Information Partners (ESIP), International Association for Social Science Information Services & Technology (IASSIST), and Research Data Alliance (RDA) to acquire and share expertise. While participants did not directly discuss visibility of data work, several data center and library professionals mentioned the challenge of getting scientists and other stakeholders to value data management.

6.3 Summary of Learning Research Data Expertise

This chapter profiles the strategies and conditions impacting learning at my two research sites. The findings document the social and individual nature of workplace learning. The case results present a variety of learning processes to acquire, share, and retain expertise that organizations can utilize in the development of data expertise. This section summarizes the key findings.

A set of workplace learning strategies for building data expertise and staffing were identified: 1) data expertise acquisition, 2) sharing, and 3) retention, similar to typologies reported in the organizational learning literature (Argote, 2011, 2013; Huber, 1991). Data expertise acquisition served to increase the knowledge and skills of professionals and groups to meet the challenges of data management and new work demands. Data expertise sharing activities focused on distributing data knowledge and skills among data workers and across organizational units. A final approach to learning was data expertise retention, embedding data expertise into shared practices and documentation for long-term preservation. These learning processes were complementary, enhancing the deep expertise and competence in employees and across the organization, but were sequential meaning that an organization must have expertise before it can focus on sharing and retention efforts. At the research sites, the learning activities blended individual and collective learning to cultivate data staff with the expertise and competence to keep pace with data management trends.

Additionally, a set of conditions was identified that can inhibit or enhance learning and organizational change: 1) sphere of influence, 2) local data community of practice, and 3) visibility of data work. The importance of influence for data professionals was illustrated, highlighting how data expertise or practice changes need to occur at multiple levels across the organization. The formation of a local data-focused community of practice enabled social learning and a shared language, worldview, practices, and systems. The visibility of data work draws attention to the importance of data expertise and work, motivating professional development of data staff. These conditions have the potential to become roadblocks or catalysts for data professional development and service innovations. The next chapter contextualizes and

integrates my study results and draws attention to study implications and future research directions.

CHAPTER 7: DISCUSSION AND CONCLUSION

Preparing to offer research data management services entails an understanding of the positions, roles, and expertise related to data work. The research in this dissertation aims to provide this understanding by investigating the research questions:

1. How do organizations develop (and support) data expertise?
    a. What roles and skills emerge from this process?
2. Why do data services and staffing develop differently in each case?

The chapter starts with a summary of the key findings and case differences followed by implications for the data professions and future research directions.

7.1 Summary of Key Findings

The research findings advance the discussion of research data expertise requirements. Many studies have contributed typologies of knowledge and skill areas required for data professionals (See Appendix E for a compilation), but the primary results are often a list of skills specific to a domain or type of data work. This study makes a substantial contribution by identifying the multiple areas and levels of expertise required and strategies for building data expertise into organizations. My study identified two organizational models for research data expertise and services. The models of NCAR and Purdue documented a set of important elements for supporting research data services that emerged from the cross-case analysis: organizational structure, boundary-spanning positions, expertise requirements, and learning strategies.

7.1.1 Organizational Structure Supporting Research Data Services

Study findings documented how organizational structure elements of teams and partners impacted the cultivation of research data expertise. In Chapter 5, the two models drew attention

to the placement of data professionals in the organizational structure. Data professionals were located in two types of teams: research teams and data-focused teams. Placing data professionals into research teams enabled data professionals to work beside the scientists and learn more about the research topic and techniques for the assigned science project. These data professionals were often isolated from other data workers in the organization limiting data expertise acquisition and sharing. In contrast, the formation of data-focused teams placed strong data management and curation expertise into a group located outside of the domain departments or labs. These data professionals were able to build an internal community for learning and sharing lessons learned and best practices. While these data teams offered a deep knowledge of data management and curation, often the expertise was not specific to research practices or data formats for a sub-discipline. Service partners providing additional expertise or services supplemented the expertise in the data and research teams. The use of partners influenced the types of expertise that a team or organization needed to develop internally, as evident in the Purdue case. Purdue librarians relied on campus partners to contribute advanced software engineering, supercomputing, and research administration knowledge and services.

7.1.2 Boundary Spanning Data Positions

Boundary spanning was identified as a distinctive feature of data positions in my study. Previous research has focused on how bridging work emphasizes connecting different organizational units and professions (Friedman & Podolny, 1992; Tushman, 1977). My analysis extends this concept to data management and curation work. Data professionals were individuals who in a similar way were bridging across disciplines or professions (e.g., earth sciences, engineering, information science) and fostering communication across these boundaries inside and outside the organization. Data positions were comprised of multiple roles with names

reflective of some of the boundaries they were bridging. For instance, data professionals were bridging the work and boundaries of engineering and science and often assuming data science and engineering responsibilities. This study result has important implications for data educators who must prepare and mentor students to design data services that bridge disciplinary and professional boundaries and foster communication across multiple boundaries.

Chapter 5 articulated a set of data roles from the cross-case analysis: Data Curator/Manager, Data Engineer, Data Liaison/Consultant, Data Scientist, and Data Service Manager. These findings were consistent with the roles of Data Curator, Scientist, and Engineer identified in several papers (Cox & Corrall, 2013; Interagency Working Group on Digital Data, 2009; Lyon, Mattern, Acker, & Langmead, 2015; Maatta, 2013; Manyika et al., 2011; Pryor & Donnelly, 2009; Sierra, 2012; Swan & Brown, 2008). Additionally, my analysis identified two additional roles of Data Liaison/Consultant and Data Service Manager that data professionals in data centers and libraries also fulfill. As seen in previous study of data curation work in digital humanities centers, data roles and responsibilities are distributed across existing staff members. In the Purdue case, data liaison roles and activities extended into other kinds of librarian positions in other library units. This study makes evident the multiple roles that data professionals play in data services and how data work bridges multiple disciplines and professions in the modern research enterprise.

7.1.3 Data Professionalization Disconnects & Dilemmas

Previous research has focused on how organizational learning initiatives are impacted by disconnects and dilemmas (Cutcher-Gershenfeld & Ford, 2005). My study extends these concepts to data professionalization innovations. The disconnects represent a range of concerns such as value of data work, data professional identities and careers, and staff resistance to

change. Underpinning these disconnects were two dilemmas - organizational priorities and organizational governance. These disconnects and dilemmas highlight the challenges in data expertise and services development and offer a possible explanation for the slow growth in data services in the academic library community (Tenopir, 2014). These findings expand on a recent survey of academic libraries that found cultivation of expertise and hiring staff as the biggest challenges for research data services (Hudson-Vitale et al., 2017) by drawing attention to the workplace challenges of integrating data professionals into established work systems, values, and career paths in an organization. The study contributes a set of strategies to harness the dilemmas. In particular, staff training on data management and curation needs to be broader than professionals working within data services. The organizational mission and strategic planning must address responsibility for data management and curation and identify objectives and goals for data services. The case results make evident that research institutions and administrators need to embrace and learn from professionalization disconnects and dilemmas in order to build sustainable data services for their communities.

7.1.4 Expertise Areas and Levels for Data Work

The set of data expertise categories improves upon previous findings by illuminating the expertise required for several types of data work. My Chapter 5 findings identified 18 types of research data expertise and configurations of expertise types and levels. The set of research data expertise types provides a description of the knowledge, skills, and experiences needed to perform high quality data management and curation (see Table 5.3 for expertise types and descriptions). The 7 expertise categories were curation, engineering, services, analytics, leadership, research, and data. In general, the findings of data expertise categories and areas are in line with other studies of data knowledge and skill requirements (Engelhardt, Strathmann, &

Mccadden, 2012; Hedstrom et al., 2015; Lee, 2009; Lyon et al., 2015; Nelson, 2016). See Appendix E for a compilation of existing data competency typologies. My findings provided a more comprehensive understanding of expertise areas required for data work. My study enhanced our understanding of previous findings on service-oriented competencies (Cox, Verbaan, & Sen, 2014; Engelhardt et al., 2012; Hedstrom et al., 2015; Lee, 2009; Lyon et al., 2015; Tammaro, Madrid, & Casarosa, 2013) by drawing attention to the need for skills in community relationship-building, collaboration, and data metrics to assess service effectiveness. This research documents the important contributions of data curators in standardizing practices, workflows, and systems.

Multiple professions are contributing knowledge to data work, demonstrating Abbott's (1988) concept of unsettled jurisdictions. NCAR and Purdue data professionals articulated expertise spanning domain sciences, engineering, information science, statistics, business, among others. Most data professionals started with a background in one area like domain science, engineering, or library science. Through experiential, observational and other acquisition strategies, data professionals cultivated knowledge and practices from other professional areas needed to perform data tasks. *Knowing-how practices* (Orlikowski, 2002) facilitated a collective expertise and competence required for data professionals to work across several professional boundaries (i.e., domain, computer, information sciences). At the time of my data collection, participants did not hold data education degrees. In the last decade, data education programs have appeared in library science, computer science, business, and other disciplines, further confirming the unsettled jurisdiction of data work and claims of expertise. Abbot's work on professions highlighted the importance of competition between professions. My study design

focusing on organizations does not enable me to comment on the claims and disputes among professions for jurisdictional control of data work.

Furthermore, my study elucidated the combination and levels of expertise needed to perform data management and curation work. Data professionals in this study were found to have built deep expertise in multiple areas throughout the course of their careers. This finding moves the discussion of data professionals' skill set away from the focus on one deep area of expertise (Bloom, 2017; Stanton et al., 2012) to skill sets with multiple areas of depth. Data work requires professionals with multiple kinds of expertise similar to contributing and interactional expertise in the 'periodic table of expertise' (Collins & Evans, 2007). Most Liaison Librarians developed an interactional expertise of data management, where they learned the language around research data enabling them to participate in consultations or conversations with scientists from multiple domains. Many Purdue Data Specialists and Curators and NCAR data professionals cultivated a contributing level of expertise, where they can perform data tasks with competence. Experienced data professionals are adept at navigating the everyday data challenges that require them to bring a deep expertise while at the same time necessitate a multi-dimensional expertise for the variety of tasks that they are expected to perform.

7.1.5 Relationship Between Data Expertise & Spheres of Influence

An important element in data service models was cultivating expertise levels of data professionals to ensure they had the appropriate sphere of influence in the organization. My study elucidated the three spheres of influence of data professionals clarifying the types of institutional actors and agency needed to support effective data services across the organization (Garud, Hardy, & Maguire, 2007; Maguire et al., 2004). The influence and agency represented the data professional's ability to develop and modify data practices, norms, and expertise at

organization, unit, and project levels. Data professionals that exhibited a larger sphere of influence than I expected based on their position were deemed as experts by their supervisors and colleagues in the areas of data management and curation. In particular, Purdue library professionals cultivated research data expertise, as legitimate peripheral participants do (Lave & Wenger, 1991), and this expertise earned them respect and influence. The library professionals engaged in the early years of data services were embedded in research projects to learn firsthand about data management. By performing data activities, they learned the language of data management and gained an understanding of the challenges and important concepts. This participation deepened their knowledge and transformed them into key players in Purdue data services. By being a central player in data services, these librarians gained support from their colleagues and agency from their supervisors.

The spheres of influence highlighted the different levels of research data expertise that need to be cultivated in libraries and data centers. For instance, the professionals with an organizational sphere of influence had an interactional level of data expertise (Collins & Evans, 2007) in order to communicate with administrators, scientists, and other stakeholders about data services and needs. However, data professionals possessing the unit sphere of influence were managers that moved from entry-level positions, where they cultivated contributing expertise, but had nurtured an interactional level of data expertise. These dual levels of expertise enabled these professionals to work with staff across the organization to ensure their lab or library adopted quality data services, practices, and systems. The data professionals with project-level influence had a contributing level of data expertise (Collins & Evans, 2007), enabling them to contribute substantially to the data activities and solutions within their teams. This comparison makes evident that organizations need to consider which levels of research data expertise (i.e.,

interactional and contributing) are needed to garner appropriate spheres of influence for their data professionals.

7.1.6 Building Data Expertise into Organizations

Local data communities of practice in my cases fostered the building, sharing, and retaining of data expertise as well as the lifelong learning of data professionals. These communities confirmed the social dimensions of expertise, fostered by performing the work and from interactions with similar professionals (Collins & Evans, 2007; Lave & Wenger, 1991; Orlikowski, 2002). These communities fostered collective learning and competence among data workers and the creation of a shared view of data, norms, practices, and systems. Shared data projects allowed opportunities for data professionals to interact and share best practices and experiences. In these communities, data professionals were valuable resources for each other to learn data expertise areas and cultivate deeper expertise in certain areas. Through communities of practice, data professionals built, exchanged, and refined the expertise areas and levels required to perform data management in their organization and respond to technological or environmental changes.

The visibility of data work in the research teams and organizations was a condition for data staffing and expertise development in the research sites. Chapter 6 draws attention to the importance of data work visibility in the research teams and organizations. By shedding light on the complexity in data processing and curation work, my work revealed the contributions of data professionals to science. The recognition of contributions fostered respect for data professionals and their expertise in their workplace as well as collegial support for data initiatives. The evidence suggested that the visibility of data practices, systems, and expertise was increasing in

both of the research sites. The visibility and respect for data work will be important factor for data centers and libraries to consider in the developing and offering data services.

Ensuring innovative and successful data services requires research institutions to carefully plan where data professionals are located in the organization, how data positions are structured, what expertise types and levels are needed, and how to support the continuing professional development of data workers. My work highlights the organizational challenges in preparing to offer research data services and offers a possible explanation for why academic libraries are slow to offer these services, a point made evident in Tenopir (2014). Research data services are defining a new space and roles for librarians on campus. With the Purdue case, librarians had to cultivate an expertise with multiple dimensions and levels to contribute to data management activities and interact with researchers. The staff training on data management also needed to be extended to other librarians across the organization. Additionally, Purdue made visible how they contribute to data management to researchers and garnered the sphere of influence appropriate for their data professionals. Learning disconnects and dilemmas provide valuable fodder for libraries to grow and nurture sustainable data services.

7.2 Case Differences in Building Data Expertise

The prominent differences between the two research sites were the 1) types of teams and 2) backgrounds of data professionals. The team structure differed by where data professionals were placed and by the reliance on service partners to provide additional expertise. As previously mentioned, NCAR placed data professionals into both science and data teams. Data professionals placed into science teams were often the one data expert, serving several data roles and requiring deep knowledge in multiple expertise areas. A second team structure was data team comprised of data positions combining the roles of data curator, liaison, and engineer and requiring a deep

expertise in the areas of curation, service, and engineering. The composition of teams impacted the creation of data positions, responsibilities, and required expertise. Science teams assigned all data roles and responsibilities to one or two positions, embedding multiple expertise areas into these positions. Data teams allowed the specialization of data roles and responsibilities across several data positions, requiring deep expertise in a few areas. Research institution administrators and managers must consider how team structure and composition impact where data positions and expertise types need to reside in the organization.

The other difference between my research sites was the professional background of data workers. NCAR data professionals had backgrounds primarily in the geosciences, but a few were from engineering with some geoscience coursework or work experiences. In contrast, Purdue data professionals often held a library and information science degree with a few having an additional domain background (e.g., social science, earth science, engineering). In general, the data professional's background impacted the types of expertise that these workers and organizations needed to develop to effectively support data services. While most NCAR data professionals were expanding their knowledge primarily in curatorial and engineering areas, Purdue data professionals focused on learning about data and research practices in multiple disciplines. The best expertise acquisition approach for curatorial gaps is observational or experiential learning; for engineering gaps it tends to be experiential or books/classroom methods, with experiential or observational strategies favored for research and data gaps. Managers could benefit from assessment tools to monitor and identify expertise gaps in their staff.

7.3 Implications for Cultivating Research Data Professions

The cultivation of the expertise and jurisdiction of data professionals will be important to

building capacity that can meet the demands of emergent research data management trends. The findings from this research on expertise has implications for 1) building data services teams and positions, 2) designing data curation education programs, and 3) supporting the data professions.

Chapter 5 presents two models for designing organizational structures to support data services and identifies a set of 4 common elements– teams, collaborations, boundary spanning positions, and roles. The use of science teams and data teams represent the different approaches to the placement of data professionals and their expertise in an organization. The use of partnerships and collaborations can address expertise and resources gaps. The set of data roles can be used to inform the design of new data service positions and to assign specific data roles to new and existing staff positions as appropriate. Managers can tailor the combination of data roles to meet their organizational goals and ensure different types of data work and responsibilities are represented in teams. The data expertise categories can inform staff development and recruitment efforts.

Chapter 5 findings illuminated how data services depend on knowledge spanning multiple professions – domain science, engineering, information science, among others. Managers can identify where these forms of expertise currently reside in the organization and identify gaps in knowledge, where organizational learning efforts are warranted. The similarities of data roles and expertise categories suggest areas for alignment in data positions and teams.

Guiding priorities for research institution and library administrators and managers in designing data services and staffing are:

- Cultivating local communities of practice for research data management to foster individual and collective learning and expertise sharing among staff.

- Creating an effective mix of data expertise areas and contributory and interactive expertise types to ensure effective engagement with user communities.
- Documenting data service innovations, decisions, and lessons learned to foster knowledge articulation and thwart knowledge loss.
- Promoting the visibility of data work and roles to highlight the contributions of data professionals to the research enterprise.
- Building positions and career ladders for data professionals, allowing for different types of agency and spheres of influence to maintain innovative data practices and continual learning.
- Documenting professionalization and learning disconnects to identify underlying dilemmas in data service innovations.
- Disseminating data training programs to data professionals and other workers to build a culture that values data work.
- Designing strategic plans and policies that incorporate research data management initiatives.
- Developing tool(s) to monitor and assess expertise gaps and to recruit new data workers to fill these voids.
- Fostering data service partnerships to enhance services and leverage external expertise and resources.

As research data expertise and service initiatives continue to mature, an important next step is further investigation of the features of successful and unsuccessful data services.

Data management and curation education programs can be informed by the Chapter 5 findings on data expertise areas. Several data education programs have based their curriculum in

the knowledge of one profession like information science, computer science, business, or domain science. The set of data expertise categories emphasizes that data professionals need expertise from multiple professions. A promising approach to data management education would be an interdisciplinary approach including courses from multiple departments to better prepare students for the variety of responsibilities and roles.

The study findings have implications for the evolving data professions and communities of practices. As previously noted, the development of a data community of practice at each research site indicates a deep appreciation for learning and improving practices and services among data professionals. As the data professions evolve, these communities may be important avenues for initiation into the profession, knowledge sharing, and advancing a shared set of competencies, theories, and best practices. These various local data communities of practice together with recently emerged data professional groups (e.g., Earth Science Information Partners, Research Data Alliance) can foster unifying efforts, such as online learning collaboratories, workshops, or joint working groups, that enable expertise sharing and innovation across these group boundaries, expanding the learning network for research data management.

7.4 Future Research

The central research problem is how organizations can build and sustain research data expertise and services. This dissertation lays the groundwork for future research into models of data expertise, data positions and their sphere of influence, and data communities of practice. Additional case studies are warranted to understand the variety of models for building research data expertise in an organization. These studies will build on my dissertation research by targeting a wide range of organizations with less mature data services and different organizational types (e.g., insurance corporations, natural history museums, field stations in

ecology, supercomputing centers, domain repositories). These new case studies will enable me to compare my NCAR and Purdue findings to other settings, resulting in a more nuanced and functional model of interest to a broader audience. This enhanced model will include more organizational approaches, elements, and recommendations for supporting research data management, laying the groundwork for recommended practices, tools, and resources to assess, analyze, and build data expertise into organizations that are planning or offering data services.

Agency and influence emerged in my research as important issues for data professionals. A second research trajectory is to further investigate where agency is embedded in research institutions and what types of agency and influence data professionals need in their positions. I propose a series of workplace studies of research and data centers in various domains and different sectors (e.g., government agency, non-profit, private firm). These studies would investigate which positions and types of professionals express agency and what level of influence they have on data services. By looking at different domains and sectors, the study would provide more comprehensive analysis of data positions and needed agency and influence. These insights will contribute to the organizational science literature on institutions and agency as well as be useful for research administrators and managers.

A final research direction is exploring communities of practice for research data management to further understand how data expertise, learning, and the data profession are evolving. Based on my findings in Chapter 6, a data community of practice in organizations provides opportunities for knowledge transfer among members. Future investigation will be aimed at understanding how these communities are shaping data expertise and role definitions and the features of successful and unsuccessful data communities of practice. This work will

inform professional development opportunities for data professionals. The insights from these future studies would draw more attention to research data expertise, work, and professionals.

7.5 Concluding Remarks

High quality data management and curation services depend on a well-trained staff with the right expertise and roles to support research. This research examined how two organizations with mature research data services developed their own data expertise and staffing, comparing approaches in a geoscience data center and an academic library. The project provided insights into the process of building research data expertise and organizational structure to support these new services. Key products were the sets of data roles and expertise categories that align to support research data services. Previously, the work and contributions of data professionals has not been adequately acknowledged in the research enterprise. My study draws attention to the contributions that data professionals make to research projects and the unique expertise they bring to science and scholarship.

Research institutions can cultivate data expertise through the set of learning strategies. To succeed, research and data centers need to increase the visibility of data workers and their contributions and to incorporate data management into strategic plans; libraries need to cultivate expertise across their staff, strategically planning which staff need which areas or levels of expertise. Attention to data expertise development will ensure a staff that can steward valuable data assets and support the demands of data-intensive research.

REFERENCES

See Appendix E for additional references specific to data competences.

Abbott, A. (1988). *The System of Professions: An Essay on the Division of Expert Labor*. Chicago, IL, USA: University of Chicago Press.

Abbott, A. (1998). Professionalism and the future of librarianship. *Library Trends*, *46*(3), 430–443.

Altman, M., & Crabtree, J. (2011). Using the SafeArchive System : TRAC-Based auditing of LOCKSS. In *Archiving 2011 Final Program and Proceedings* (pp. 165–170). Retrieved from http://www.imaging.org/IST/store/epub.cfm?abstrid=44591%5Cnhttp://www.box.net/shared/8py6vl9kxivo6u21rkn8

American Council of Learned Societies. (2006). *Our Cultural Commonwealth*. Retrieved from http://www.acls.org/cyberinfrastructure/ourculturalcommonwealth.pdf

Anderson, R. G., Greene, W. H., McCullough, B. D., & Vinod, H. D. (2005). *The Role of Data & Program Code Archives in the Future of Economic Research*. St. Louis, MO, USA.

Argote, L. (2011). Organizational learning research: Past, present and future. *Management Learning*, *42*(4), 439–446. http://doi.org/10.1177/1350507611408217

Argote, L. (2013). *Organizational Learning: Creating, Retaining and Transferring Knowledge* (2nd ed.). New York, New York, USA: Springer Science+Business Media. http://doi.org/10.1007/978-1-4614-5251-5

Association of Research Libraries. (2006). *To Stand the Test of Time: Long-term Stewarship of Digital Data Sets in Science and Engineering*. Retrieved from http://files.eric.ed.gov/fulltext/ED528649.pdf

Atkins, D. (2003). *Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. Retrieved from http://www.nsf.gov/cise/sci/reports/atkins.pdf

Bachrach, C. A., & King, R. B. (2004). Data sharing and duplication: Is there a problem? *Archives of Pediatric and Adolescent Medicine*, *158*, 931–932.

Baker, K. S., Mayernik, M. S., Thompson, C. A., Nienhouse, E., Williams, S., & Worley, S. (2015). Envisioning and enacting a coherent organization-wide view of data. In *Proceedings of the International Digital Curation Conference.* Retrieved from http://hdl.handle.net/2142/73150

Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Englewood Cliffs, N.J.: Prentice-Hall.

Bates, M. J. (1999). The invisible substrate of information science. *Journal of the American Society for Information Science*, *50*(12), 1043–1050.

Bermes, E., & Fauduet, L. (2011). The human face of digital preservation: Organizational and staff challenges, and initiatives at the Bibliotheque nationale de France. *International Journal on Digital Curation*, *6*(1), 226–237.

Bloom, J. (2017). Astronomy. In *Roundtable on Data Science Post-Secondary Education Meeting #2: Examining the Intersection of Domain Expertise and Data Science*. Irvine, CA. Retrieved from http://sites.nationalacademies.org/DEPS/BMSA/DEPS_176954

Board on Earth Sciences and Resources. (2002). *Geoscience Data and Collections: National Resources in Peril*. Washington, DC, USA.

Board on Life Sciences. (2003). *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*. Washington, DC, USA.

Bonn, G. S. (1959). Training for activity in acientific documentation work. In *Proceedings of the International Conference on Scientific Information* (pp. 1441–1487). National Academies Press. Retrieved from http://www.nap.edu/catalog/10866.html

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. http://doi.org/10.1007/s10606-012-9169-z

Botticelli, P., Fulton, B., Pearce-Moses, R., Szuter, C., & Watters, P. (2011). Educating digital curators: Challenges and opportunities. *International Journal of Digital Curation*, *6*(2), 146–164.

Brandt, D. S. (2013). Purdue University Research Repository: Collaborations in data management. In J. Ray (Ed.), *Research Data Management: Practical Strategies for Information Professionals* (pp. 325–346). Ashland, OH, USA: Purdue University Press.

Briggs, C. L. (1986). *Learning How to Ask: A Sociolinguistic Appraisal of the Role of the Interview in Social Science Research*. Cambridge, UK: Cambridge University Press.

Brown, E. (2010). I know what you researched last summer: How academic librarians are supporting researchers in the management of data curation. *The New Zealand Library & Information Management Journal*, *52*(1), 55–69.

Bryant, R., Lavoie, B., & Malpas, C. (2017). *A Tour of the Research Data Management (RDM) Service Space. The Realities of Research Data Management, Part 1.* Dublin, OH, USA. Retrieved from www.oclc.org/research

Burnett, K. M., & Bonnici, L. J. (2006). Contested terrain: Accreditation and the future of the profession of librarianship. *The Library Quarterly*, *76*(2), 193–219.

Burwell, S. M., Vanroekel, S., Park, T., & Mancini, D. J. (2013). *Memorandum for the Heads of*

Executive Departments and Agencies: Open Data Policy - Managing Information as an Asset. Washington, DC, USA. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf

Carnegie Foundation for the Advancement of Teaching. (2015). Standard Listings - Carnegie Classification of Institutions of Higher Education. Retrieved March 20, 2017, from http://carnegieclassifications.iu.edu/lookup/srp.php?clq=%7B%22basic2005_ids%22%3A%2215%22%7D&start_page=standard.php&backurl=standard.php&limit=0,50

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics*, *3*(3), 191–209.

Choo, M. (2014). *Exploring Knowing in Practice An Ethnographic Study of Teams in the Agile Setting*. Linnaeus University.

Choudhury, S. G. (2008). Case study in data curation at Johns Hopkins University. *Library Trends*, *57*(2), 211–220. http://doi.org/10.1353/lib.0.0028

Choudhury, S. G., Palmer, C. L., Baker, K. S., & Dilauro, T. (2013). Levels of services and curation for high functioning data. In *2013 International Digital Curation Conference*.

Cohan, L., & Craven, K. (1961). *Scientific Information Personnel: The New Profession of Information Combining Science, Librarianship and Foreign Language*. New York, New York, USA: Modern Language Associaton of America.

Collins, H., & Evans, R. (2002). The third wave of science studies: Studies of expertise and experience. *Social Studies of Science*, *32*(2), 235–296. http://doi.org/10.1177/0306312702032002003

Collins, H., & Evans, R. (2007). *Rethinking Expertise*. Chicago, IL, USA: The University of Chicago Press.

Córdoba, J.-R., Pilkington, A., & Bernroider, E. W. N. (2012). Information systems as a discipline in the making: comparing EJIS and MISQ between 1995 and 2008. *European Journal of Information Systems*, *21*(5), 479–495. http://doi.org/10.1057/ejis.2011.58

Corrall, S., Keenan, M. A., & Afzal, W. (2013). Bibliometrics and research datamanagement services: Emerging trends in library support for research. *Library Trends*, *61*(3), 636–674. Retrieved from http://dx.doi.org/10.1353/lib.2013.0005

Council on Library and Information Resources. (2008). *No Brief Candle: Reconceiving Research Libraries for the 21st Century*. Retrieved from http://www.clir.org/pubs/reports/pub142/pub142.pdf

Cox, A. M., & Corrall, S. (2013). Evolving academic library specialties. *Journal of the American Society for Information Science and Technology*, *64*(8), 1526–1542. http://doi.org/10.1002/asi

Cox, A., Verbaan, E., & Sen, B. (2014). A spider, an octopus, or an animal just coming into existence? Designing a curriculum for librarians to support research data management. *Journal of eScience Librarianship*, *3*(1). http://doi.org/10.7191/jeslib.2014.1055

Cragin, M. H., Chao, T. C., & Palmer, C. L. (2011). Units of evidence for analyzing subdisciplinary difference in data practice studies. In *JCDL'11* (pp. 441–442). Ottawa, Ontario, Canada. http://doi.org/10.1145/1255175.125522

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *368*(1926), 4023–38. http://doi.org/10.1098/rsta.2010.0165

Crane, D. (1972). *Invisible colleges: Diffusion of Knowledge in Scientific Communities*. Chicago, IL, USA: University of Chicago Press.

Creamer, A., Morales, M. E., Crespo, J., Kafel, D., & Martin, E. R. (2012). An assessment of needed competencies to promote the data curation and management librarianship of health sciences and science and technology librarians in New England. *Journal of eScience Librarianship*, *1*(1), 18–26.

Creswell, J. W. (2009). *Research Design: Qualitative, Quantiative, and Mixed Methods Approaches* (3rd ed.). Thousand Oaks, CA, USA: SAGE Publications, Ltd.

Currall, J., Johnson, C., & McKinney, P. (2007). The world is all grown digital…How shall a man persuade management what to do in such times. *International Journal of Digital Curation*, *2*(1), 12–28.

Cutcher-Gershenfeld, J., & Ford, J. K. (2005). *Valuable Disconnects in Organizational Learning Systems*. New York, NY: Oxford University Press.

Dasler, R., Muñoz, T., & Nilsen, K. (2013). Beyond the repository : Rethinking data services at the University of Maryland. In *Special Libraries Association Annual Conference 2013*. Alexandria, VA, USA: Special Libraries Association.

Day, M. (2008). Toward distributed infrastructures for digital preservation: The roles of collaboration and trust. *International Journal of Digital Curation*, *3*(1), 15–28.

Digital Curation Centre. (2008). The DCC Curation Lifecycle Model. Retrieved June 6, 2017, from http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf

Dougherty, D., & Dunne, D. D. (2012). Digital science and knowledge boundaries in complex innovation. *Organization Science*, *23*(5), 1467–1484. http://doi.org/http://dx.doi.org/10.1287/orsc.1110.0700

Downs, R. R., & Chen, R. S. (2010). Self-assessment of a long-term archive for interdisciplinary scientific data as a trustworthy digital repository. *Journal of Digital Information*, *11*(1).

Edmonson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, *44*(4), 350–383.

Emerson, R. M., Fretz, R. I., & Shaw, L. L. (2011). *Writing Ethnographic Fieldnotes*. Chicago, IL, USA: University of Chicago Press.

Engelhardt, C., Strathmann, S., & Mccadden, K. (2012). *DigCurv: Report and Analysis of the Survey of Training Needs.* Retrieved from http://www.digcur-education.org/eng/Resources/Report-andanalysis- on-the-training-needs-survey

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363–406.

Fisher, K. E., Durrance, J. C., & Hinton, M. B. (2004). Information grounds and the use of need-based services by immigrants in Queens, New York: A context-based, outcome evaluation approach. *Journal of the American Society for Information Science and Technology*, *55*(8), 754–766.

Flanders, J., & Hamlin, S. (2013). TAPAS: Building a TEI publishing and repository service. *Journal of the Text Encoding Initiative*, (5), 1–9. http://doi.org/10.4000/jtei.788

Freese, J. (2007). Replication standards for quantitative social science. *Sociological Methods and Research*, *32*(2), 153–172.

Friedman, R. A., & Podolny, J. (1992). Differentiation of boundary spanning roles: Labor negotiations and implications for role conflict. *Administrative Science Quarterly*, *37*(1), 28–47.

Garud, R., Hardy, C., & Maguire, S. (2007). Institutional entrepreneurship as embedded agency :

An introduction to the special issue. *Organizational Studies*, *28*(7), 957–969. http://doi.org/10.1177/0170840607078958

Garud, R., Jain, S., & Kumaraswamy, A. (2002). Institutional entrepeneurship in the sponsorship of common technological standards: The case of Sun Microsystems and Java. *Academy of Management Journal*, *45*(1), 196–214.

Gil, Y., Chan, M., Gomez, B., & Caron, B. (2014). *EarthCube: Past, Present, and Future*. Retrieved from http://earthcube.org/document/2014/earthcube-?-past-?-present-?-future

Gladwell, M. (2008). *Outliers: The Story of Success*. New York, New York, USA: Little, Brown and Co.

Gold, A. K. (2007). Cyberinfrastructure, data, and libraries, part 2: Libraries and the data challenge: Roles and actions for libraries. *D-Lib Magazine*, *13*(9/10).

Greenberg, J., White, H. C., Carrier, S., & Scherle, R. (2009). A metadata best practice for a scientific data repository. *Journal of Library Metadata*, *9*(3–4), 194–212.

Gutmann, M. P., Abrahamson, M., Adams, M. O., Altman, M., Arms, C., Bollen, K., Carlson, M., Crabtree, J., Donakowski, D., King, G., Lyle, J., Maynard, M., Pienta, A., Rockwell, R., Timms-Fierra, L., & Young, C. H. (2009). From preserving the past to preserving the future: The Data-PASS project and the challenges of preserving digital social science data. *Library Trends*, *57*(3), 315–337. http://doi.org/10.1353/lib.0.0039

Hedges, M., Haft, M., & Knight, G. (2012). FISHNet: Encouraging data sharing and reuse in the freshwater science community. *Journal Of Digital Information*, *13*. Retrieved from http://journals.tdl.org/jodi/article/view/5884

Hedstrom, M., Dirks, L., Fox, P., Goodchild, M., Joseph, H., Larsen, R., Palmer, C. L., Ruggles, S., Schindel, D., & Wandner, S. (2015). *Preparing the Workforce for Digital Curation*.

Washington, DC. Retrieved from https://www.nap.edu/catalog/18590/preparing-the-workforce-for-digital-curation

Hey, T., Tansley, S., & Tolle, K. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Fourth+Paradigm#4

Hoffman, G. L. (2012). Could the functional future of bibliographic control change cataloging work? An exploration using Abbott. *Journal of Library Metadata*, *12*(2–3), 111–126. http://doi.org/10.1080/19386389.2012.699825

Holdren, J. P. (2013). *Memorandum for the Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research*. Washington, DC, USA. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Holdren, J. P. (2014). *Memorandum for the Heads of Executive Departments and Agencies: Improving the Management of and Access to Scientific Collections*. Washington, DC, USA. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_memo_scientific_collections_march_2014.pdf

Hou, C., Betancourt, T., & Mayernik, M. (2015). Crediting a climate model dataset like a movie? A case study in data attribution. In *Proceedings of International Digital Curation Conference*. Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/IDCC15/175_Creatingaclimatemodel.pdf

Hou, C., Thompson, C. A., & Palmer, C. L. (2014). Profiling open digital repositories in the atmospheric and climate sciences : An initial survey. In *Proceedings of the 77th ASIS&T Annual Meeting*. Seattle, WA, USA.

Huber, G. (1991). Organizational learning : The contributing processes and the literatures. *Organization Science*, *2*(1), 88–115.

Hudson-Vitale, C., Imker, H., Johnston, L. R., Carlson, J., Kozlowski, W., Olendorf, R., & Stewart, C. (2017). *SPEC Kit 354: Data Curation*. Washington, DC. Retrieved from http://publications.arl.org/Data-Curation-SPEC-Kit-354/

Interagency Working Group on Digital Data. (2009). *Harnessing the Power of Digital Data for Science and Society*. Retrieved from http://www.nitrd.gov/About/Harnessing_Power.aspx

Jenne, R. (1975). *Data Sets for Meterological Research. NCAR Tech Note.* Boulder, CO. Retrieved from http://n2t.net/ark:/85065/d7s46rbh

Jenne, R. (2005). *An Oral History with Roy Jenne/Interviewer: Stuart (Bill) Leslie.* Boulder, CO. Retrieved from http://n2t.net/ark:/85065/d7513wj3

Johnston, L. R. (2017). *Curating Research Data: A Handbook of Current Practice* (Volume 2). Chicago, IL, USA: Association of College and Research LIbraries.

Kim, Y., Addom, B. K., & Stanton, J. M. (2011). Education for eScience professionals: Integrating data curation and cyberinfrastructure. *International Journal of Digital Curation*, *6*(1), 125–138.

Kimberly, J. R., & Evanisko, M. J. (1981). Organizational innovation: The influence of individual, organizational, and contextual factors on hospital adoption of technological and administrative innovations. *Academy of Management Journal*, *24*(4), 689–713.

Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Hock, H., Lautenschlager, M., Schindler, U., Sens, I., & Wächter, J. (2006). Data publication in the open access initiative. *Data Science Journal*, *5*, 79–83.

Kolb, D. (1984). *Experiential Learning: Experience as the Source of Learning and Development*. Englewood Cliffs, N.J.: Prentice-Hall.

Kuhn, T. S. (1996). *The Structure of Scientific Revolutions* (3rd Editio). Chicago: The University of Chicago Press.

Kuipers, T., & van der Hoeven, J. (2009). *Insights into Digital Preservation of Research Output in Europe: PARSE-Insight Survey Report*. Retrieved from http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

Kurzweil, R. (2004). The Law of Accelerating Returns. *Alan Turing: Life and Legacy of a Great Thinker*, 1–50. Retrieved from http://link.springer.com/chapter/10.1007/978-3-662-05642-4_16

Kvale, S. (1996). *InterViews: An Introduction to Qualitative Research Interviewing*. Thousand Oaks, CA, USA: SAGE Publications, Ltd.

Latour, B. (1999). Circulating reference: Sampling the soil in the Amazon forest. In *Pandora's Hope: Essays on the Reality of Science Studies* (pp. 24–79). Cambridge, MA, USA: Harvard University Press.

Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. New York, NY: Cambridge University Press.

Lee, C. A. (2009). *Functions and Skills: Dimension 2 of Matrix of Digital Curation Knowledge and Competencies (Version 18)*. Retrieved from https://ils.unc.edu/digccurr/digccurr-functions.html

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic Inquiry*. Beverly Hills, CA, USA: SAGE Publications, Ltd.

Lord, P., & Macdonald, A. (2003). *Data curation for e-Science in the UK: An Audit to Establish Requirements for Future Curation and Provision.* Retrieved from http://www.jisc.ac.uk/uploaded_documents/e-scienceReportFinal.pdf

Lyon, L., Mattern, E., Langmead, A., & Acker, A. (2015). Applying translational principles to data science curriculum development. In *iPres 2015*. Chapel Hill, North Carolina. Retrieved from http://d-scholarship.pitt.edu/id/eprint/27159

Lyon, L., Wright, S., Corti, L., Edmunds, S., & Bennett, F. (2013). What is a data scientist? In *2013 International Digital Curation Conference*. Amsterdam, Netherlands: Digital Curation Centre.

Maatta, S. L. (2013). The emerging databrarian. *Library Journal*, *138*(17), 26.

Maguire, S., Hardy, C., & Lawrence, T. B. (2004). Institutional entrepreneurship in emerging fields: HIV/AIDS treatment advocacy in Canada. *The Academy of Management Journal*, *47*(5), 657–679. http://doi.org/10.2307/20159610

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big Data: The Next Frontier for Innovation, Competition, and Productivity*.

Mayernik, M. S., Daniels, M. D., Ginger, K. M., Kelly, K. M., Marlino, M. R., Williams, S. F., & Wright, M. J. (2012). *Data Citations within NCAR / UCP*. Boulder, CO. Retrieved from http://dx.doi.org/10.5065/D6ZC80VN

Mayernik, M. S., Davis, L., Kelly, K., Dattore, B., Strand, G., Worley, S. J., & Marlino, M. (2014). Research center insights into data curation education and curriculum. In Ł. Bolikowski, V. Casarosa, P. Manghi, P. Goodale, N. Houssos, & J. Schirrwagen (Eds.),

Theory and Practice of Digital Libraries-TPDL 2013 Selected Workshops, Communications in Computer and Information Science (Vol. 416 CCIS, pp. 239–248). Switzerland: Sprinter International Publishing. http://doi.org/10.1007/978-3-319-08425-1

Morris, S. (2013). *Issues in the Appraisal and Selection of Geospatial Data*. Retrieved from http://hdl.loc.gov/loc.gdc/lcpub.2013655112.1

National Aeronautics and Space Administration. (2015). EOSDIS DAACs. Retrieved April 21, 2015, from https://earthdata.nasa.gov/about-eosdis/science-system-description/eosdis-components/eosdis-daacs

National Institutes of Health. (2003). Final NIH Statement on Sharing Research Data (NOT-OD-03-032). Retrieved from http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

National Science Foundation. (2015). Dissemination and Sharing of Research Results. Retrieved from http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Nature Publishing Group. (2009). Guide to Publication Policies of the Nature Journals. Retrieved from http://www.nature.com/authors/editorial_policies/availability.html

Nelson, M. S. (2016). *Scaffolding for Data Management Skills: From Undergraduate Education through Postgraduate Training and Beyond*. West Lafayette, IN.

Nicolini, D. (2011). Practice as the site of knowing: Insights from the field of telemedicine. *Organization Science*, *22*(3), 602–620. http://doi.org/10.1287/orsc.1100.0556

Onwebguzie, A. J., & Leech, N. L. (2005). The role of sampling in qualitative research. *Academic Exchange Quarterly*, *9*(3), 280–284.

Orlikowski, W. J. (2002). Knowing in practice: Enacting a collective capability in distributed organizing. *Organization Science*, *13*(3), 249–273.

http://doi.org/10.1287/orsc.13.3.249.2776

Palmer, C. L., & Cragin, M. H. (2008). Scholarship and disciplinary practices. *Annual Review of Information Science and Technology*, *42*(1), 163–212. http://doi.org/10.1002/aris.2008.1440420112

Palmer, C. L., Thompson, C. A., Baker, K. S., & Senseney, M. (2014). Meeting data workforce needs: Indicators based on recent data curation placements. In *iConference 2014 Proceedings*. Berlin, Germany.

Palmer, C. L., Thompson, C. A., Tenopir, C., Allard, S., Mayernik, M. S., & Kreft, J. (2014). Responding to emerging data workforce demand: Harnessing data center expertise. In A. Grove (Ed.), *Proceedings of the 77th ASIS&T Annual Meeting, vol. 51*.

Parsons, M. A., & Berman, F. (2013). The Research Data Alliance: Implementing the Ttchnology, practice and connections of a data infrastructure. *Bulletin of the American Society for Information Science and Technology*, *August/Sep*, 33–36.

Patton, M. Q. (1999). Enhancing the quality and credibility of qualitative analysis. *Health Services Research*, *34*(5 Pt 2), 1189–1208.

Postle, B. R., Shapiro, L. A., & Biesanz, J. C. (2002). On having one's data shared. *Journal of Cognitive Neuroscience*, *14*(6), 838–40. http://doi.org/10.1162/089892902760191063

Pryor, G., & Donnelly, M. (2009). Skilling up to do data: Whose role, whose responsibility, whose career ? *International Journal of Digital Curation*, *4*(2), 158–170.

Purdue University Libraries. (2016). *Libraries Fact Sheet*. West Lafayette, IN. Retrieved from https://www.lib.purdue.edu/sites/default/files/admin/current_fact_sheet.pdf

Raven, B. H., Schwarzwald, J., & Koslowsky, M. (1998). Conceptualizing and measuring a power/interaction model of interpersonal influence. *Journal of Applied Social Psychology*,

28(4), 307–332. http://doi.org/10.1111/j.1559-1816.1998.tb01708.x

Ray, J. M. (2013). *Research Data Management: Practical Strategies for Information Professionals*. Ashland, OH, USA: Purdue University Press.

Rusbridge, C. (2007). Create, curate, re-use: the expanding life course of digital research data. In *EDUCAUSE Australasia 2007 Advancing Knowledge Pushing Boundaries. CAUDIT*. Retrieved from http://hdl.handle.net/1842/1731

Savolainen, R. (2009). Epistemic work and knowing in practice as conceptualizations of information use. *Information Research*, *14*(1), 1–16.

Schneider, B. (2004). Building a scientific community: The need for replication. *Teachers College Record*, *106*, 1471–83.

Sierra, T. (2012). Staffing for the future research questions: ARL university lIbrary hiring in 2011. In *ARl Fall Forum 2012*.

Simon, H. A., & Chase, W. G. (1973). Skill in chess. *American Scientist*, *61*, 394–403.

Simons, H. (2009). *Case Study Research in Practice*. London, UK: SAGE Publications, Ltd.

Stake, R. E. (1995). *The Art of Case Study Research*. Thousand Oaks, CA, USA: SAGE Publications, Ltd.

Stanton, J., Palmer, C. L., Blake, C., & Allard, S. (2012). Interdisciplinary data science education. In N. Xiao & L. R. McEwen (Eds.), *Special Issues in Data Management* (pp. 97–113). Washington, DC: American Chemical Society. http://doi.org/10.1021/bk-2012-1110.ch006

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. http://doi.org/10.1023/A:1008651105359

Steinhart, G. (2011). DataStaR: A data sharing and publication infrastructure to support research. *Agricultural Information Worldwide*, *4*(1), 16–20.

Stodden, V., Guo, P., & Ma, Z. (2013). Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLOS ONE*, *8*(6), e67111. http://doi.org/10.1371/journal.pone.0067111

Sustein, B. S., & Chiseri-Strater, E. (2012). *Fieldworking: Reading and Writing Research*. Boston, MA: Bedford/St. Martin's.

Swan, A., & Brown, S. (2008). *The Skills, Role and Career Structures of Data Scientists and Curators: An Assessment of Current Practice and Future Needs*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.147.8960&rep=rep1&type=pdf

Tammaro, A. M., Madrid, M., & Casarosa, V. (2013). Digital curators ' education: Professional identity vs. convergence of LAM (Libraries, Archives , Museums). In M. Agosti, F. Esposito, S. Ferilli, & N. Ferro (Eds.), *IRCDL 2012. Communications in Computer and Information Science* (pp. 184–194). Bari, Italy: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-35834-0_19

Taylor, C. C. W. (1991). *Plato's Protagoras*. Oxford, UK: Clarendon Press.

TEKSystems. (2013). Trends IT industry big data … The next frontier. *IT Industry Trends*, pp. 1–5.

Tenopir, C. (2014). Research data services: What opportunities exist in academic libraries? In *In Proceedings of the ASIST Annual Meeting (Vol. 51)*. John Wiley and Sons Inc. http://doi.org/10.1002/meet.2014.14505101026

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), e21101. http://doi.org/10.1371/journal.pone.0021101

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. http://doi.org/10.1016/j.lisr.2013.11.003

Thompson, C. A., Mayernik, M. S., Palmer, C. L., Allard, S., & Tenopir, C. (2015). LIS programs and data centers: Integrating expertise. In *iConference 2015 Proceedings*.

Thompson, C. A., & Palmer, C. L. (2014). Lessons learned from job placements and employer interviews. In *In Proceedings of the ASIST Annual Meeting (Vol. 51)*. John Wiley and Sons Inc. http://doi.org/10.1002/meet.2014.14505101026

Traweek, S. (1988). *Beamtimes and Lifetimes: The World of High Energy Physicists*. Cambridge, MA, USA: Harvard University Press.

Trosow, S. E. (2001). Jurisdictional disputes and the unauthorized practice of law. *Legal Reference Services Quarterly*, *20*(4), 1–18.

Tushman, M. L. (1977). Special boundary roles in the innovation process. *Administrative Science Quarterly*, *22*(4), 587–605. Retrieved from http://www.jstor.org/stable/2392402

University Corporation for Atmospheric Research. (2017a). About NCAR. Retrieved from https://ncar.ucar.edu/about-ncar

University Corporation for Atmospheric Research. (2017b). Climate & Global Dynamics Laboratory – About. Retrieved from https://www2.cgd.ucar.edu/about/our-mission

Varvel, V. E. J., Palmer, C. L., Chao, T., & Sacchi, S. (2011). *Report from the Research Data Workforce Summit*. Champaign, IL, USA. Retrieved from http://hdl.handle.net/2142/25830

Vaughan, D. (1999). The role of the organization in the production of techno-scientific knowledge. *Social Studies of Science*, *29*(6), 913–943.

Veinot, T. C. (2007). The eyes of the power company: Workplace information practices of a vault inspector. *The Library Quarterly*, *77*(2), 157–180.

Wagner, M. (2007). On the relationship between environmental management, environmental innovation and patenting: Evidence from German manufacturing firms. *Research Policy*, *36*(1587–1602).

Wenger, E. (1998). *Communities of Practice: Learning, Meaning, and Identity*. Cambridge, UK: Cambridge University Press.

Whitbeck, C. (2005). The responsible collection, retention, sharing, and interpretation of data. *Online Ethics Center for Engineering and Science*. Retrieved from http://onlineethics.org/reseth/mod/data.html

Witt, M. (2012). Databib: An online bibliography of research data repositories. In *Proceedings of ACRL Digital Curation Interest Group*. Anaheim, California: ALA Annual Conference. Retrieved from http://www.ala.org/lita/sites/ala.org.lita/files/content/conferences/forum/2012/PosterSessionDescriptions.pdf

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing data curation profiles. *International Journal of Digital Curation*, *4*(3), 93–103. http://doi.org/10.2218/ijdc.v4i3.117

Yin, R. K. (1977). Production efficiency versus bureaucratic self-interest: Two innovative processes? *Policy Sciences*, *8*(4), 381–399.

Zahedi, M., & Babar, M. A. (2014). Towards an understanding of enabling process knowing in

global software development : A case study. In *Proceedings of the 2014 International Conference on Software and System Process (ICSSP'14)* (pp. 30–39). http://doi.org/10.1145/2600821.2600836

Zimmerman, A. S. (2007). Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, *7*(1–2), 5–16. http://doi.org/10.1007/s00799-007-0015-8

Zimmerman, A. S. (2008). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science, Technology & Human Values*, *33*(5), 631–652. http://doi.org/10.1177/0162243907306704

APPENDIX A: INTERVIEW SCHEDULES OF QUESTIONS

NCAR Case Study
Data Professional Interview

Goals:

- Roles for data management
- Data services offered
- Required data expertise for this lab's work
- Relationship of data expertise to services
- History of data efforts if possible

I'd like to learn a bit about this lab.

1. How does the work of this lab relate to the larger organization?
   - Please describe the lab and the work that happens here.
   - If on a team or work group, what is its role?

2. What services are offered in terms of scientific data management?
   - PROBE: What is data for this lab?
   - What are the data services or tools that you are most proud of here?
   - Which area do you think has the most support in this lab?

3. Who has responsibility for scientific data management in this lab?
   - What are the typical position titles and duties? Could you share with me any job descriptions or ads?
   - Number of positions?
   - How many employees do data management exclusively vs. part of their job?

I'm interested in learning more about the history of data efforts in this lab.

4. Please tell me the story of how data efforts started in this lab.
   a. When did it start? What was the initial driver or motivation?
   b. How have these services evolved over time in this lab? Any other important drivers or pivotal events?
   c. What have been the barriers to these data efforts?
   d. History of data roles and positions in this lab? When was the first person hired? How have roles and responsibilities changed over time here?

I'd like to learn a bit about your role.

5. What is your role within your lab?
   - Please describe a typical day for you.
   - What is your position title and your key duties?
   - Do you do data management exclusively vs. part of your job?

- Do you interact with other data professionals in this lab?

6. What does "data stewardship and engineering" mean to you?
   - With which professions and/or fields, do you associate your work?
   - Do you identify as a data professional? Why or why not?
   - Do you belong to any professional organizations, societies, or other groups?

7. What do you consider to be the expertise of data managers (or data scientists)?
   - What contributions do data professionals bring to this lab/project?
   - How is this different from what IT or scientists offer?

8. What core skills and knowledge do new hires require in order to perform a job like yours?
   - What background do you need? How important is this background?
   - Is prior experience required?
   - How did you learn to do this work?

9. For the data services that you offer, which skills are needed for each service?
   - Are there any skills that all services require?
   - Any skills unique to only one data service/task?

In closing, I'd like to discuss data positions.

10. Since you've been in the field, how have data roles and positions changed?
    - What was the field like when you first began? What is it like now?
    - How do you see these roles changing in the next 5 years? 10 years?
    - Any advice for a person wishing to enter scientific data management?

11. May we contact you for follow up or to clarify on your answers?

12. Can you recommend anyone else that we should talk to about these issues?

Thank you for participating in the interview! For the purpose of describing the participant pool in reports, I have a few demographic questions.

13. What is the title of your current position? _____

14. How long have you been employed in your current position? _____years _____months

Supplementary Geoscience Data Centers
DCERC Schedule of Questions

Interview Goals:

- To understand the employer's data workforce roles and needs
- To understand what kinds of data positions exist (duties, responsibilities) and where they fit within the organization
- To assess how well the DCERC program is addressing employers' needs for data curation professionals

FOR ORGANIZATIONS THAT HAVE HIRED DCERC GRADUATE:

Now, we are interested in the graduate of our Data Curation Education in Research Centers (DCERC) program that you hired. We would like to ask you a few questions about this employee.

- Can you please describe [Fellow Name] role within the organization?
  - Job title?
  - Primary responsibilities?
  - To whom does s/he report?
  - Who does s/he supervise?

- In regard to the hiring process, what made [Fellow Name] competitive within the pool of applicants?
  - How did the NCAR internships contribute to their competitiveness?

- In your opinion, what would you change or add to her/his preparation?
  - Which area(s) was the graduate well-prepared for?

FOR ALL EMPLOYERS:

The following questions are part of a larger study of data and research centers and understanding their staffing needs in terms of data management.

We have a few quick questions about your data operations.

1. Can you give me an overview of the data operations in your department and how it relates to the rest of the organization?
   a. Who produces the data?
   b. How many departments are involved?
   c. Who is the service community or are the users of these data?

I'd like to learn about your staff for data management/operations.

2. Who has responsibility for scientific data management?

- What are the typical position titles and duties? Could you share with me any job descriptions or ads?
- Number of positions?
- How many employees do data management exclusively vs. part of their job?

3. How well is your staff addressing the data needs of the organization?
    - Which needs are addressed well? Which needs are not? Why?
    - What do you see as the strengths and weaknesses of your staff?
    - What would you change about their preparation if you could?

I'd like to focus specifically on the core knowledge and skills of your data management staff.

4. What kinds of skills does a person need to do data management in your organization? Please describe in detail.
    - What background do your staff have? How important is it for them to have this background?
    - Is prior experience required?
    - What other qualities do you look for when hiring new data professionals?

5. How do you *find* the right kind of person for data management? Please explain.
    - How do you *retain* data professionals? Please explain.
    - How many data professionals do you anticipate needing over the next 5 years? 10 years?

6. Is there a career path for data professionals in your organization?
    - Please give me an example of someone who made a career in data management at this organization.
    - Is their experience typical?

7. *To what extent have you or your staff participated in training related to data management?
    - What types of training? Topics?
    - If money were no object, what other training would you like your staff?

8. What do you see as the challenges in preparing data professionals?
    - How will this change in the next 5 years? 10 years?

9. Key pieces of the DCERC program are the data curation curricula in topics such as data management, information organization, metadata, and then the summer internship at the National Center for Atmospheric Research where our students get hands-on experience working with scientific data and scientists. We are interested in expanding opportunities for students. Do you think your organization could serve as an internship or field experience site for data curation students?
    - What do you think these internship opportunities could look like?

10. In closing, what do you think will be one or two of the biggest challenges for your data operations in the future?
    - Have you seen the challenges change over time?
    - Do you believe that your challenges are similar or different from the challenges of your peer institutions?

11. Are there any final thoughts you have about preparing data professionals from your perspective?

12. Would you be willing to do a follow-up interview?

Thank you for participating in the interview! For the purpose of describing the participant pool in reports, I have a few demographic questions.

1. What is the title of your current position? _____

2. How long have you been employed in your current position? _____years _____months

Purdue Case Study
Librarian Interview

Goals:

- History and future of data management efforts in the library
- Drivers or pivotal events for these efforts
- Data services offered
- Roles for data management

Let's start by talking about the role of the university library.

1. What is the mission of the library?
   - Please describe the work that happens here.
   - How does the work relate to the larger university?

2. What services are offered in terms of research data?
   - What is data?
   - What are the data services or tools that you are most proud of here?
   - Which area do you think has the most support

I'm interested in learning more about the history of data efforts at the library.

3. Please tell me the story of how data efforts started in the library.
   - When did it start? What was the initial driver or motivation?
   - How have these services evolved over time? Any other important drivers or pivotal events?
   - What have been the barriers to these data efforts?
   - History of data roles and positions? When was the first person hired? How have roles and responsibilities changed over time here?

I'd like to learn a bit about your role.

4. What is your role within the library?
   - What is your position title?
   - Please describe a typical day. Key duties?
   - Do you do data management exclusively vs. part of your job?
   - How frequently do you interact with users?
   - How did you learn to do this work?

5. Do you identify as a data professional? Why or why not?
   - With which professions and/or fields, do you associate your work?
   - Do you belong to any professional organizations, societies, or other groups?

I'd like to switch gear and discuss your staff for data services/operations.

6. Who has responsibility for scientific data management in the library?
    - What are the typical position titles and duties? Could you share with me any job descriptions or ads?
    - Number of positions?
    - How many employees do data management exclusively vs. part of their job?

7. What do you consider to be the expertise of data librarians or curators?
    - What contributions do data professionals bring to research?
    - How is this different from what IT or researchers offer?

8. How have data roles and positions changed over time?
    - How do you see these roles changing in the next 5 years? 10 years?

9. In closing, what do academic libraries need to do to stay relevant to the research community?
    - What should their role in the community be?
    - Is your library on the leading edge, in the middle of the pack or behind the times in terms of staying relevant?

10. May we contact you for follow up or to clarify on your answers?

11. Can you recommend anyone else that we should talk to about these issues?

Thank you for participating in the interview! For the purpose of describing the participant pool in reports, I have a few demographic questions.

12. What is the title of your current position? _____

13. How long have you been employed in your current position? _____years _____months

Supplementary Academic Libraries
RDA Fellowship Interview Questions

Prepare for interview:

- Review library's organizational chart
- Pull library demographics from the website
- Confirm signed consent form
- Prepare follow-up questionnaire and instructions for interviewee

I'd like to learn about how your library is organized to support research data management. I thought that it would be useful to use your library's organizational chart as a tool for discussing this.

1. First, I would like to confirm that I understand correctly from the organizational chart the placement of research data management services in your library. [DESCRIBE WHAT I SAW IN THEIR ORG CHART]
    - Are there informal relationships that exist to support RDM not represented in your organization chart?
    - What would you say works well for your organization in supporting RDM? Please provide examples.
    - What would you say doesn't work quite so well? Please provide examples.

Organizational structure also represents the flow of responsibility and accountability. I would now like to talk about who has responsibility for RDM in your library.

2. Currently, who has responsibility for research data in the library? Responsible means the person that performs the task.
    - What are the position titles and responsibilities? How many have the term, *data*, in their position title?
    - Do they work on RDMS exclusively or as part of their job?
    - Which unit(s) are involved in RDMS? How do the units work together?

3. Who is accountable for research data services in the library? Accountable means the person who must answer for the correctness or completion of the task.
    - What are the position title(s)?
    - Is the position accountable for all RDMS activities or a portion? Please explain.

4. What has to be consulted within the library about operations supporting RDM? Consulted means those whose opinions are consulted; there is a two-way communication about the task.
    - What are the position title(s)?
    - How are decisions about RDM services made in this library?

5. Who has to be informed about RDM services in the library? Informed means those who must be kept up-to-date on the progress of the task, so it's more one-way communication.
        - Who is responsible for ensuring that communication happens?

6. Do you partner with any campus or external units to offer data services? If so, please describe.
    - How did these partnerships get started?
    - What are the services or programs that you collaborate on?
    - How are data service responsibilities split between the partners?
    - Is this a formal or informal relationship? Are services paid for from the library or partners?

7. Could you tell me the history of the library's structure and staffing for research data services?
    - When did it start? What were the motivations?
    - Any important pivotal events or champions for RDMS?
    - When was the first RDMS person appointed/hired? Was a team formed?
    - How have organizational approaches changed over time?

8. Do you think that your library's structure and staffing for RDMS is stable? Please explain why.
    - How do you expect RDMS structure to change in the future?
    - What have been the barriers or facilitators? Could you give me an example of how they impacted the structure/staffing?

I'd like to switch gears and focus a bit on the research data services.

9. What programs or services do you offer specifically for research data now?
    - What is data?
    - Who is the community being served?
    - How do you engage users? Service points?
    - Are there any data-related services that fall outside of the research data services?
    - Where does the funding for RDMS come from?

10. What core skills and knowledge do new librarians require in order to work in data services?
    - What background do they need? How important is this background?
    - Is prior experience required?

11. In closing, what do academic libraries need to do to stay relevant to the research community?
    - What should their role in the community be?
    - Is your library on the leading edge, in the middle of the pack, or behind the times in terms of staying relevant?

12. Finally, is there anything else that you would like to tell us about your data services or organizational structure?

13. If I have additional questions, would you be willing to do a follow-up interview?

Thank you for your time! I have a few demographic questions to describe the sample of participants that I spoke with for this study.

1. What is the title of your current position? _____

2. How long have you been employed in your current position? _____ years

3. Is your library part of an institution that grants:
   - ☐ Doctoral degrees
   - ☐ Master's degrees
   - ☐ Bachelor's degrees
   - ☐ Associate degrees

4. What is the number of librarians in your organization? _____

APPENDIX B: HUMAN SUBJECT REVIEW BOARD APPROVALS

UIUC Institutional Review Board

UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

Office of the Vice Chancellor for Research

Office for the Protection of Research Subjects
528 East Green Street
Suite 203
Champaign, IL 61820

August 4, 2015

Janet Wyatt Eke
ASST DIR, CIRSS
Library & Information Science
344 Lis
501 E Daniel
Champaign, IL 61820

RE: *Data Curation Education in Research Centers (DCERC)*
IRB Protocol Number: 12051

EXPIRATION DATE: August 3, 2018

Dear Ms. Eke, Ms. Thompson and Dr. Tenopir:

Thank you for submitting the completed IRB application form for the amendment and request for a three year extension of exemption approval for your project entitled <*Data Curation Education in Research Centers (DCERC)*>. Your project was assigned Institutional Review Board (IRB) Protocol Number 12051 and reviewed. It has been determined that the research activities described in this application meet the criteria for exemption at 45CFR46.101(b)(2).

Related to Dr. Tenopir's role as a co-investigator, please supply a copy of the University of Tennessee IRB approval letter or correspondence from their office that indicated that they would like to rely on the UIUC IRB to be the IRB of record for this exempt research.

This determination of exemption only applies to the research study as submitted. Please note that additional modifications to your project need to be submitted to the IRB for review and exemption determination or approval before the modifications are initiated.

We appreciate your conscientious adherence to the requirements of human subjects research. If you have any questions about the IRB process, or if you need assistance at any time, please feel free to contact me at the OPRS office, or visit our website at http://www.irb.illinois.edu.

Sincerely,

Ronald A Banks

Ronald Banks, MS, CIP
Human Subjects Research Coordinator, Office for the Protection of Research Subjects

NCAR Review Board

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Dr. Michael J. Thompson
NCAR CHIEF OPERATING OFFICER
P. O. Box 3000, Boulder, CO 80307-3000 USA
Phone: 303.497.1500 • Fax: 303.497.1194
mjt@ucar.edu • www.ncar.ucar.edu

MEMORANDUM

To: Matthew Mayernik

From: UCAR Human Subjects Committee M.J. Thompson

OHRP IRB Number: IRB00006222 (U Corp. for Atmospheric Research)
Assurance Number: FWA00012567

Date: July 9, 2015

Subject: Human Subjects Committee Review for study "Data Expertise and Service Development Study"
HSC Memo #2015-11

Review: Exempt

Category: 2 under 45 CFR Part 690

The Human Subjects Committee (HSC) has reviewed the study protocol "Data Expertise and Service Development Study" and finds that it is exempt under 45 CFR 690.101, category 2 because research is involving survey procedures where there is no information obtained or recorded that can identify the human subjects in any way.

You are required to update the HSC at any time if any of the aspects of your study that involve human subjects changes. The UCAR HSC's approval of this study is for human subjects research purposes only and in no way reflects any management opinion about the study or its potential results.

We appreciate your conscientious adherence to the requirements of human subjects research. If you have any questions about this process, or if you need assistance at any time, please feel free to contact the HSC (hsc@ucar.edu) or visit our website at http://www.ucar.edu/hsc/.

The National Center for Atmospheric Research is operated by the University Corporation for Atmospheric Research under sponsorship of the National Science Foundation

APPENDIX C: QUALITATIVE CODEBOOK

Code Group	Code Name	Definition
Expertise	Knowledge and skills	The expertise, knowledge and skills needed to perform RDM
Expertise	Learning	The avenues for learning to do RDM work such as peer-to-peer, courses, books, etc.
Expertise	Staffing background/ expertise	The RDM staff background or unique expertise or contributions they bring to the work
RDM Staff	RDMS staff/structure history	The history of RDM structure and staffing such as new position of data curator or when teams formed or dissolved
RDM Staff	RDMS structure	The organizational approach to support RDM such as units, teams, embedded staff, solo positions
RDM Staff	RDM positions	The positions, roles of staff in RDMS
RDM Staff	RDM partnership	Units or organizations that partner on RDM and their roles
RDM Staff	RDM staffing size	The number of RDM staff
RDM Staff	RDM responsibility	The specific responsibilities of RDM staff, performing the RDM work day to day
RDM Staff	RDM staff strengths	Strengths of the RDMS staff
RDM Staff	RDM staff weakness	RDM staff weaknesses or additional needs
RDM Staff	RDM staffing trends	Future trends, growth, needs, or changes in RDM staffing
RDM Staff	Career narratives	Stories of RDM staff careers or how they fell into RDM
RDM Service	RDM service history	History of RDM services
RDM Service	RDM services	The activities related to RDM such as operations, programs, or services
RDM Service	Barriers	The barriers or challenges to offering RDM services
RDM Service	Facilitators	The facilitators to offering RDM services
RDM Service	Funding	The funding source(s) for the RDM services and staff; mentions of funding agencies or budget allocations
RDM Service	RDM service trends	Anticipated trends or changes in RDMS services
RDM Service	Data practices	How the work of RDM is performed or carried out, including the social aspects and artifacts used
RDM Service	User community	The community or audience served by the organization, demographics on users, users' needs
Data Profession	Professional activities	Professional activities, conferences, journals, and associations that RDM participate in
Data Profession	Professional claims	A claim of specialized knowledge over an area of work
Data Profession	Professional identities	How professionals identify or which groups they affiliate their work

Table C.1. Table of qualitative codes and definitions

Table C.1 (cont.)

Code Group	Code Name	Definition
Data Profession	Professional jurisdiction	The control over a service or area of work, jurisdiction boundaries, and competition with other professions
Organization	Org History	The history of the organization
Organization	Org Mission	The mission, goals and purpose of the organization
Organization	Org Structure	The departments, units, and teams and their relationship in the organization
Organization	Org culture	The organizational culture or climate such as service-oriented, flat hierarchy, siloes
Organization	Decision-making	The process for making decisions in the organization
Organization	RDM support	The support or buy-in for RDM services
Organization	Local Data Policies	Mentions of local policies related to research data
Organization	Stakeholders	Stakeholders related to the organization such as funding agencies, publishers, government (non-RDM staff and partners)
Research	Domain area	The primary domain or research area of the organization
Research	Instruments & Methods	Mentions of research instruments, models, or methods
Research	Data and products	The data and products produced, managed, and/or preserved by the organization such as types and holdings
Misc	Environmental factors	Factors external to the organization and unit that impact RDM
Misc	Invisible work	RDM Work, expertise and skills that are unnoticed by others, including work that is routinized so that it fades into background or not given the respect, legitimacy or valued as other work
Misc	Scalability	Mentions related to scaling RDM services
Misc	Tensions	Tensions or frictions related to RDM - two forces that resist or clash such as budget allotment versus actual needs, scientific vs LIS expertise, research vs service perspective, services offered vs. advertised
Misc	Value of RDM staff	The perception or value of RDM staff by others within or outside of organization
Misc	Archetype Characteristics	Features of org structure and RDM services that cause us to assign the archetype
Misc	Ethics	Mention of ethical issues in RDM services such as security, privacy, copyright or licensing.

APPENDIX D: RESEARCH DATA EXPERTISE CASE RESULTS

This appendix provides the case comparison of the research data expertise types and categories to supplement Chapter 5 results. The report is organized by expertise categories reporting NCAR results and then Purdue results.

Data

NCAR

Across NCAR's history, the distinct data types of expertise have been essential to data work. These types were strongly emphasized by participants as important for data professionals. Participants in all three teams drew attention to the importance of Data Handling expertise (Type #1 in Table 5.3). An NCAR-Climate Data Engineer colorfully highlighted his/her data processing skills as "…you give me a file and I can turn it into a well-constructed, well-designed NetCDF file. Then, I can take any NetCDF file and beat it to a bloody pulp to make it do exactly what I want it to" (NCAR 206). Data handling skills were a requirement for new data professionals being hired. A NCAR-Solar Data Scientist described the desired qualifications as: "if you've had experience handling data at any level…any processing, have you done any validating" (NCAR 207). Data handling expertise was a foundation for data work.

At NCAR, data work demanded that data professionals be familiar with current and emerging trends such as best practices, funding or publishing requirements, open science stakeholders, and even domain, national, and international trends (Type #2 in Table 5.3). Several interviewees had participated in data-related committees and/or conferences at discipline, university, agency, national, or international levels, keeping abreast of emerging trends.

Purdue

Purdue librarians have been improving their expertise in Data types. Early on Purdue embedded librarians into science teams and conducted exploratory research projects, such as the *Investigating Data Curation Profiles across Research Domains* project funded by the Institute of Museum and Library Services. These projects allowed librarians to learn about handling research data and the diversity of data types and structures (Type #1 in Table 5.3). Furthermore, these early data efforts taught librarians about the landscape of research stakeholders, trends, and requirements (Type #2 in Table 5.3). The quote from the Repository Director illustrated the importance of understanding the landscape of data trends when conducting data consultations:

> If you're a librarian and you walk into a researcher's office, the researcher asks you what can you do for me. 'Ok, I'm familiar with the funder requirements for data management in your discipline. I can help you identify an appropriate data repository for you to submit your data to. I'm familiar with publishers in your field and understand the author requirements for data deposit or supplementary data. I can come in and speak with your graduate students about effective data management practices…I can help connect you to tools that can make all these things easier RE3 data, DMPTool, PURR…' (Purdue 101)

These Data types of expertise emerged as important knowledge and skill areas for data librarians. In two recent job advertisements for data librarians, experience handling data and research outputs and knowledge of data trends were listed as preferred requirements.

Research

NCAR

Across the three NCAR teams, all interviewees had responsibilities associated with research support, needing the research types of expertise. Data professionals had prior experience in a research setting and had participated in different activities – planning, collection, and dissemination. These experiences provided first-hand knowledge of the research process and its activities and needs (Type #3 in Table 5.3). As the NCAR-Archive Data Service Manager illustrated: "They [data professionals] have to understand science, how science is done, what do these measurements mean, what do these numbers mean" (NCAR 211). As an NCAR-Climate Data Engineer noted the problem in hiring staff without a research background to do this work:

> It's important in the science field to have people that have some science background…If you historically look at NASA, they are always trying to get people to use their data. If you look at some of the products they put out, my view is that they had a bunch of…non-science people put stuff into the HDF funnel and it came out with a bunch of stuff. It made it very difficult to use the [NASA] data. (NCAR 205)

The intimate knowledge of the research process enabled data professionals to design services and data products that met the needs of scientists.

At NCAR, knowledge of common research instruments and models was essential for data professionals (Type #4 in Table 5.3). While the data professionals in observational teams had knowledge of instruments and their history, the data professionals placed in weather and climate modeling teams had extensive knowledge and experience with community models in their sub-discipline. For instance, an NCAR-Solar Data Engineer working in a solar modeling group reported the importance of her/his understanding the "principles of modeling" and history of solar community models that s/he brings to data work (NCAR 208). Across the eight NCAR-Archive Data Engineers, the team has knowledge of both observational instruments and computational simulations used in the geosciences. Similar to the previous categories, the knowledge of the research process, instruments, and simulation techniques were strongly emphasized as important for data professionals working at NCAR.

Purdue

Purdue staff cultivated expertise in research practices and workflows. Purdue librarians were embedded in research teams, providing an opportunity for them to learn more about the research process (Type #3 in table 5.3). These experiences gave librarians an understanding of the research workflows, practices, challenges, and terminology. As one liaison librarian describes:

> It's knowing the domain language and jargon is really important for data work…and asking lots and lots of questions to clarify what that terminology means in that context because the same jargon can be used in different disciplines in a different context meaning different things so you have to be willing to understand what the differences are. (Purdue 104)

Historically, librarians prided themselves on understanding their user communities and needs. The extension of library services to research data has placed emphasis on learning more about the research process.

At Purdue, the categories of research instruments and models expertise (Type #4 in Table 5.3) were not observed in the interview transcripts. The library serves a multi-disciplinary audience, using a large variety of research instruments and techniques. The Purdue model for data services includes relying on service partners for additional expertise that the data professionals do not possess.

Curation

NCAR

At NCAR, the three teams exhibited different levels of emphasis for the curation types of expertise. All participants reported that they depended on expertise in the organizing principles of open access, description, and discovery (Type #5 in Table 5.3). Their knowledge and experiences with metadata generation, identifiers, and foundational retrieval systems were valuable contributions to their work and teams. One NCAR-Climate Data Engineer described his/her contribution to climate modeling work as, "I understand things like provenance, ontologies, metadata standards, and DOIs…which none of the other… engineers in the division can understand" (NCAR 206). In all three teams, data professionals emphasized the importance of information search and retrieval systems, being able to design search engines and enable connections to third party aggregators like Virtual Solar Observatory, Earth System Grid, and Global Change Master Directory.

Collection development refers to a systematic approach to organizing data and materials into thematic collections to meet the needs of user communities. NCAR-Archive staff were the only participants to emphasize their ability to build valuable, archival collections of data based on relevant themes like "a region, over a certain time series," bringing the knowledge of "how the data should be packaged and provided to users" (NCAR 202). NCAR-Archive is a distinct team of data professionals archiving a variety of data types, providing more opportunities for staff to create groupings of products, whereas NCAR-Climate and Solar data professionals are often working with one model type, instrument data, or science topic.

At NCAR, the application of shared standards developed by the research communities is increasingly considered best practice across the atmospheric and climate sciences as well as other research fields. The Standardization type of expertise refers to knowledge of research-data-related standards and of how to develop compliant practices (Type #6 in Table 5.3). All participants contributed a blended knowledge of earth science standards (e.g., Federal Geographic Data Committee, ISO 19115, Ecological Metadata Language) and multi-disciplinary standards (e.g., DataCite, DOIs) to data work. The specific standards varied by labs –for instance, NETCDF and CF standards for NCAR-Climate, FITS for NCAR-Solar, and DIFF, OAIS, and OAI-PMH for NCAR-Archive. Standardization involved translation work where all data professionals must be knowledgeable of how to translate local practices and products into standard-compliant outputs. As one NCAR-Climate Data Engineer describes participation in the Program for Climate Model Diagnosis and Inter-comparison (PCMDI):

> …all these different modeling groups, ran all these experiments…I was in charge of that whole project. Basically taking what I considered the raw model output and translating it into their [PCMDI] requirements… (NCAR 206)

The NCAR-Archive Data Service Manager highlighted how serving interdisciplinary communities required conforming to several standards, leading to a special initiative "to get our

metadata under control" (NCAR 211). A permanent Data Engineer was dedicated to understanding different metadata standards and creating a local standard that "maps into the different ISO standards" (NCAR 213).

The range of work involved in preserving data begins with planning and continues through re-appraisal and deaccession (Type #7 in Table 5.3). The NCAR-Archive data practices, in comparison to NCAR-Climate and Solar, exhibited stronger commitment to archiving and preservation principles and best practices. All NCAR-Archive Data Engineers were familiar with preservation standards and models (e.g., OAIS), archive-friendly data formats, digital preservation techniques such as migration and emulation, provenance, preservation technologies (e.g., LOCKSS), and other archival best practices. The NCAR-Archive archiving practices have evolved over time to become more standardized across data types as the data service manager described the transition from "no systematic operation" to "archive things in the same way using the same tool" (NCAR 211). At NCAR-Climate, only one Data Engineer drew attention to the importance of archival knowledge. This group often submitted their models to other archives for preservation. At NCAR-Solar, no data professionals reported archival practices or knowledge.

At NCAR, assessing quality included knowledge of best practices for data and metadata quality and for conducting quality assurance checks (Type #8 in Table 5.3). Data professionals in all three NCAR labs elaborated on the importance of data validation and quality control skills. As an NCAR-Solar Data Scientist emphasized the ability to: "make sure that the data are validated…you at least know that the numbers should be in this range… know what the parameters are that define a valid image" (NCAR 207). As NCAR serves a multi-disciplinary community of end users, the quality of metadata to support reuse was a practice present in all three labs. Data professionals possessed knowledge of metadata best practices to ensure accuracy and clarity: "But for data to be reused outside of their originating discipline you have to kind of reduce the ambiguities and make it so that people in another discipline can understand it" (NCAR 203).

At NCAR, a final curation expertise was data ethics and legal issues (Type #9 in Table 5.3). NCAR complied with mandates and requirements for data sharing from multiple sources. For instance, federal legislation restricted the dissemination of research data to specific countries. As one NCAR-Archive Data Engineer described their compliant data practices and systems:

> There are embargoed countries that are right now Cuba, Iran, and North Korea. We can't serve them. So if their [user] IP address comes from there, then we can't serve them… we can't serve data to them, and we can't even answer emails from them. (NCAR 203)

The curation types of expertise have grown in importance over time as the NCAR-Archive Data Engineer forecasts that: "…we're [data professionals] going to be a lot more tightly coupled with the library sciences, because...the data is geoscience's paper of the future" (NCAR 213). Curatorial knowledge was a priority for NCAR data professionals to develop.

Purdue

Curation expertise was a strong common theme across the responses from Purdue participants. Traditionally, librarians and archivists have been organizers and preservers of information. Purdue extended their traditional expertise in organizing collections, preserving, assessing quality, and standardizing approaches (Types #5,6,7,8 in Table 5.3) to research data.

The librarians described how they learned how to "apply the principles of library science to these [data] problems" and translate LIS concepts to a new object resulting in "collection development and data, reference and data, classification, organization, description…of data" (Purdue 101). Deeper levels of expertise resided in certain librarian positions such as description in the Metadata Specialist, preservation in the Archivist, and domain practices or norms in Liaison Librarians.

At Purdue, the final area of curation expertise was navigating ethical and legal issues related to research data (Type 9 in Table 5.3). All data professionals and liaison librarians provided advice and training on privacy and legal issues, while repository staff encountered privacy challenges in their data work. The library leveraged additional expertise on compliance from the sponsored research office on specific funder requirements and the campus IT unit on technical aspects such as data governance and security. Since the inception of data services at Purdue, the existing curation expertise in the library has been translated to working with research data objects.

Engineering

NCAR

At NCAR, the engineering expertise (Type #10 in Table 5.3) was critical to scientific data management and preservation work. Data work required software and applications to achieve their scientific goals. NCAR has been recognized as a leader in developing earth science tools and structures, such as Earth System Grid (predecessor to Earth System Grid Federation), NCAR Command Language for analysis/visualization, and NetCDF self-describing data format. NCAR has cultivated a staff with considerable experience in utilizing a variety of software packages and scripting languages (e.g., MatLab, R, Python). A few seasoned data professionals witnessed the evolution of scientific technologies and described how they added new programming languages or applications to their repertoire of skills during their career. As one data professional in NCAR-Archive reflects on her/his career: "…we work with readers, we work with binary data, Fortran, C, Perl, Python, IDL, Matlab code…I have programmed in so many different languages in my lifetime" (NCAR 203).

At NCAR, data professionals utilized software engineering skills to design new applications when off-the-shelf software did not meet their needs. Scientific data work at NCAR relied on data professionals with the ability to understand the user requirements and envision computational solutions (existing or not existing yet); all interviewees expressed this expertise as very important for data professionals. The NCAR-Archive Data Service Manager noted how data professionals have to keep updating their technical skills over time:

> …job requirements have changed over time. And that means that the staff I have, they have to be self-taught. I mean they [staff] have to evolve over time. And this is a very important thing…the underlying capability of coding, understanding things, and putting things on the web that just keeps changing and getting better and better.

As technology advanced, harnessing technology for data solutions continued to be important over time at NCAR.

Purdue

At Purdue, the emphasis placed on technical and engineering skills (Type #10 in Table

5.3) varied by library unit. Across the library, all data-related library staff were knowledgeable about the plethora of data tools, software, and technologies. The Purdue-Archive team exhibited the most comfort with programming languages, requirement analysis, and user-interface design. The Software Engineer highlighted needing to know multiple scripting and programming languages such as "PHP in Linux, MySQL…the front and developmental language like CSS, HTML, JavaScript" (Purdue 107). The Purdue-Archive Staff emphasized the importance of requirement analysis in supporting data infrastructure. The Software Engineer summarized the critical piece as "…the transformation of the [user] requirements into the [technical] specification is very important…" (Purdue 107). While the engineering expertise emerged in the Purdue-Archive team, the library depended on the campus IT department for more extensive high performance computing (HPC), advanced software development, and systems engineering expertise.

Service

NCAR

At NCAR, understanding data use and user*s* was a common expertise, where data professionals recognized the potential uses of data sets and the needs of end users, making the connection between access and use (Type #11 in Table 5.3). One NCAR-Archive data professional summarized this expertise as the ability to "put yourself in the scientists' shoes to recognize what's needed in terms of service" (NCAR 211). Several interviewees elaborated on how this knowledge was used to design user-friendly services and applications.

Across the three NCAR teams, data work included helping users discover data sets and training users on using data portals or software (Types #12-13 in Table 5.3). User education involving workshops, blogs, videos, and tutorials depended on knowledge of teaching and instructional design. At NCAR-Archive and Climate, the user education needs have grown over time resulting in the dedication of a data professional to instruction activities.

At NCAR, building relationships with the user community type consisted of developing and cultivating relationships, requiring communication skills and trust building (Type #14 in Table 5.3). This expertise type appeared in the interviews only with NCAR-Archive staff. NCAR-Archive data professionals often attend professional conferences and workshops to meet with community members. The elevation of this expertise has spurred NCAR-Archive to dedicate, on a part-time basis, a data professional to outreach activities. Interestingly, NCAR-Solar and Climate data professionals did not mention this skill but often identified as a community member: "we don't just serve the data. We use the data ourselves. We're scientists here too" (NCAR 207). In addition to the end users, data professionals interacted and partnered with geoscience and data stakeholders to offer services (Type #15 in Table 5.3). The ability to communicate and collaborate was vital to these partnerships. At NCAR-Solar and Climate labs, data professionals often deposited data and/or model code and outputs in community data facilities or aggregators, involving interactions with data center staff. All NCAR-Archive data professionals mentioned harvesting data from other data centers (e.g., NOAA, Japanese Meteorological Agency, European Centre for Medium-Range Weather Forecasts) to develop reanalysis data sets. The service side of data work continued to respond to the data needs of the NCAR scientists and broader geoscience community.

At NCAR, data metrics involved understanding of how to measure the archive performance and the value of its collections and holdings (Type #16 in Table 5.3). At both

NCAR-Archive and NCAR-Solar, several data professionals drew attention to the importance of data metrics in their work. An NCAR-Solar Data Scientist noted an informal, local group sharing best practices for data metrics looking at "…providing data and then keeping track of who your users are, what do people use a lot. Are they [users] having trouble finding things? Are they [users] having trouble using it" (NCAR 207). While NCAR-Climate data professionals did not discuss metrics, an NCAR-Climate data scientist noted collecting "anecdotal evidence of people telling me they're using it" and the challenge of making service decisions based on little data (NCAR 108). At NCAR, the expertise area of data metrics has grown in importance with recent trends in data-intensive science and data citation initiatives.

Purdue

At Purdue, the service expertise emphasized heavily the interactions with users. Data services have been targeted at Purdue scientists, scholars, and students (Type #11 in Table 5.3). All librarian staff, both data and non-data positions, possessed an intimate knowledge of the Purdue community. Knowing the users, their practices, and needs was stressed as important for data services in order to design effective services and systems. Data professionals and repository staff worked across many domains and relied on liaison librarians and research office staff for specific knowledge about disciplines or academic departments.

Purdue liaison librarians and data professionals acknowledged the skills in helping users find data (Type #12 in Table 5.3). Librarians utilized their reference interview skills to understand a user's request and their knowledge of their holdings and retrieval systems to identify data sets. In addition to data discovery, librarians provided training for users on DMPs, data literacy, or research tools (Type #13 in Table 5.3). The data services team brought knowledge of teaching and curriculum development to these services. A valuable output has been the data literacy curriculum and competencies.

At Purdue, a central theme from all librarians was their ability to build relationships with their community (Type #14 in Table 5.3). In order to cultivate and nurture these relationships, librarians possessed trust-building, communication, and outreach skills. As the Repository Director emphasized the work required for these relationships:

> It takes so much work to get to the point where you actually get in a pick-up truck and get driven out to the field, to a hut, where there is a water-till that is collecting water to where there is a device, an instrument, that is recording data to be able to say like, 'oh ok so this is how you do your data collection.' So all those layers of awareness, of relationship building at the enterprise level…To make that collaboration really be a true collaboration to me that's the real work of it. (Purdue 101)

A Liaison Librarian described the value of cultivating these relationships:

> I think that building strong liaison ties more than just being the representative from the library, but actually building relationships is really important and not very well understood. But because I have a relationship with my faculty where we're colleagues and it's not just that I'm the librarian that serves them they are more willing to look at us as partners in something like data…having actual relationships, in-depth relationships is part of why Purdue is successful" (Purdue 103)

The recent restructuring of the Purdue-Archive staff to include a position with more outreach duties is a sign of the importance of this expertise in data service work.

At Purdue, librarians collaborated with research and data stakeholders (Type #15 in Table 5.3). Data professionals were active in research data collaborations and coordination groups like Research Data Alliance. Furthermore, in 2013 Purdue's DataBib, a catalogue of data repositories, was merged with RE3 catalogue, requiring the data librarians to collaborate with RE3 and DataCite during this transition phase. The liaison librarians also collaborated with stakeholders in their assigned communities. As a Liaison Librarian noted:

> I do a lot with agricultural data so I'm working with the National Ag Library on their new repository; it's called Ag Data Commons…I'm working with the north central experiment directors- working with their respective librarians and developing data services for this group. So I'm leading that initiative. (Purdue 103)

Services skills have been a cornerstone of traditional librarianship and have been embedded in the new data services.

As similar to peer libraries, Purdue collected user metrics to enable collection development, staffing arrangements, and program planning. Data metrics track the impact and value of research data services (Type #16 in Table 5.3). The Repository Director and Senior Data Specialist assessed research data services using measurements like number of consultations, data sets curated, and trainings offered. The data repository platform tracked additional information on its holdings – number of views, downloads, and data citations. While no librarians reported data metrics as their expertise, it was obvious that metric knowledge was present in their practices and systems.

Analytics

NCAR

All NCAR interviewees exhibited data analytic and visualization expertise (Type #17 in Table 5.3). These categories were strongly emphasized in interviews with NCAR-Solar and Climate data professionals. For instance, an NCAR-Climate Data Engineer describes how understanding evapotranspiration analyses helped to design better data services:

> We show how to access not only some of the model data, but then some other data that may not be model generated, it may be observational-based, and how that can be used to derive some of these quantities. So, here learning what objectives the people want and what data they need for it…how can we make that data available to them at least through our tools. (NCAR 205)

An NCAR-Solar Data Scientist highlighted how visualizations improve data discovery in their web portal:

> We try to provide movies of [solar] activities, pointing to other data sets and combining data just to try to point people towards things we think the scientific community is very interested in. (NCAR 207)

Purdue

While the data analytics and visualization expertise themes (Type #17 in Table 5.3) were not observed in the interview transcripts, Purdue collaborated with campus partners to provide additional expertise that is not present in the library. Two interviewees noted domain analytics expertise in other campus units.

Leadership

NCAR

Leadership was not emphasized by NCAR participants as an important skill set for data work. Since the inception of NCAR-Archive, the team has always had a manager providing oversight and coordination. In both the NCAR-Solar and Climate labs, data professionals were placed into science teams with a Principal Investigator providing overall science leadership, but data professionals exhibited leadership related to data management.

Purdue

Similar to NCAR, Purdue interviewees did not emphasize the importance of leadership as a skill set for data professionals. At Purdue, the leadership for data services fell to four positions – Dean of Libraries, Associate Dean of Research, Senior Data Specialist, and Repository Director. The Dean and Associate Dean brought leadership and ability to make connections across campus. As co-manager for research data services, the Senior Data Specialist and Repository Director possessed skills in communicating a vision, planning services and work, and staff management.

APPENDIX E: COMPILATION OF EXISTING DATA COMPETENCES

This table maps the data competences reported in the literature to the 11 data knowledge and skill areas in the *Preparing the Workforce for Digital Curation* report (Hedstrom et al., 2015).

Data Knowledge & Skill Areas	**Source**
General	Marty, 2008; Heidorn et al, 2007; Lyon et al, 2015; Mayernik et al, 2014; Lee, 2009; Engelhardt et al, 2012
Data practices	Heidorn et al, 2007; Lyon et al, 2015; Cox et al, 2014; Tammaro et al, 2012; Lee, 2009; Calzada Prado & Marzal, 2013; Nelson, 2016; Munoz & Renear, 2011
Data collection and management	Mayernik et al, 2014; Cox et al, 2014; Lee, 2009; Engelhardt et al, 2012; Calzada Prado & Marzal, 2013; Nelson, 2016; Heidorn et al, 2007; Munoz & Renear, 2011; Lyon et al, 2015
Data analytics	Nelson, 2016; Lyon et al, 2015; Calzada Prado & Marzal, 2013
Visualizations and presentation	Nelson, 2016; Lyon et al, 2015; Calzada Prado & Marzal, 2013
Technologies, tools, and interoperability	Marty, 2008; Cox et al, 2014; Tammaro et al, 2012; Lee, 2009; Engelhardt et al, 2012; Mayernik et al, 2014; Nelson, 2016; Munoz & Renear, 2011
Policy and planning	Heidorn et al, 2007; Cox et al, 2014; Tammaro et al, 2012; Lee, 2009; Engelhardt et al, 2012
Values and principles	Lee, 2009; Calzada Prado & Marzal, 2013; Nelson, 2016
Services and support	Cox et al, 2014; Lyon et al, 2015; Tammaro et al, 2012; Lee, 2009; Engelhardt et al, 2012; Calzada Prado & Marzal, 2013
Management and administration	Heidorn et al, 2007; Munoz & Renear, 2011; Lyon et al, 2015; Cox et al, 2014; Tammaro et al, 2012; Lee, 2009; Engelhardt et al, 2012
Archiving and preservation	Nelson, 2016; Tammaro et al, 2012; Lee, 2009; Engelhardt et al, 2012; Calzada Prado & Marzal, 2013; Nelson, 2016
Other competences	
Relationship building	Marty, 2008; Tammaro et al, 2012
Learn new technologies	Lyon et al, 2015
Collection development	Lyon et al, 2015; Nelson, 2016

Table E.1. Table of data competences and sources

List of data competences resources:

Calzada Prado, J., & Marzal, M. Á. (2013). Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri: International Journal of Libraries & Information Services, 63*(2), 123–134.

Cox, A. M., Verbaan, E., & Sen, B. (2014). A spider, an octopus, or an animal just coming into existence? Designing a curriculum for librarians to support research data management. *Journal of eScience Librarianship 3*(1). http://dx.doi.org/10.7191/jeslib.2014.1055

Engelhardt, C., Strathmann, S., & McCadden, K. (2012). DigCurv: Report and Analysis of the Survey of Training Needs. Retrieved from http://www.digcur-education.org/eng/Resources/Report-andanalysis-on-the-training-needs-survey

Hedstrom, M., Dirks, L., Fox, P., Goodchild, M., Joseph, H., Larsen, R., Palmer, C. L., Ruggles, S., Schindel, D., & Wandner, S. (2015). *Preparing the Workforce for Digital Curation*. Washington, DC. Retrieved from https://www.nap.edu/catalog/18590/preparing-the-workforce-for-digital-curation

Heidorn, P.B., Palmer, C.L., Cragin, M.H., & Smith, L.C. (2007). Data curation education and biological information specialists. *Proceedings of DigCCurr2007: An International Symposium on Digital Curation*, April 18-20, Chapel Hill, NC. Retrieved from http://www.ils.unc.edu/digccurr2007/papers/heidornEtal_paper_8-2.pdf

Lee, C. (2009). Matrix of Digital Curation Knowledge and Competencies Overview (Version 13). Retrieved from http://www.ils.unc.edu/digccurr/digccurr-matrix.html

Lyon, L., Mattern, E., Acker, A.,& Langmead, A. (2015) Applying translational principles to data science curriculum development. In: *iPres 2015*, November 2-6, 2015, Chapel Hill, North Carolina.

Marty, P. F. (2008). Cultural Heritage Information Professionals (CHIPs) Workshop Report. Sarasota, FL. Retrieved from http://chips.ci.fsu.edu/chips_workshop_report.pdf

Mayernik, M. S., Davis, L., Kelly, K., Dattore, B., Strand, G., Worley, S.J., & Marlino, M. (2014). Research center insights into data curation education and curriculum. In L. Bolikowski, V. Casarosa, N. Houssos, P. Manghi, & J. Schirrwagen (Eds). *Theory and Practice of Digital Libraries - TPDL 2013 Selected Workshops* (pp. 239-248). Retrieved from http://link.springer.com/chapter/10.1007%2F978-3-319-08425-1_26

Muñoz, T., & Renear, A.H. (2011). Issues in humanities data curation. *Discussion paper circulated at the Palo Alto Summit on Humanities Data Curation*, Stanford, CA, June 23, 2011. Retrieved from http://hdl.handle.net/2142/30852 and http://cirssweb.lis.illinois.edu/paloalto/whitepaper/premeeting/

Nelson, M. S. (2016). Scaffolding for data management skills: From undergraduate education through postgraduate training and beyond. *Purdue University Research Repository*. doi:10.4231/R7QJ7F9R

Tammaro, A. M., Madrid, M., & Casarosa, V. (2013). Digital curators ' education: Professional identity vs. convergence of LAM (Libraries, Archives , Museums). In M. Agosti, F. Esposito, S. Ferilli, & N. Ferro (Eds.), *IRCDL 2012. Communications in Computer and Information Science* (pp. 184–194). Bari, Italy: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-35834-0_19