Building Data Expertise into Research Institutions: Preliminary Results

Cheryl A. Thompson
Graduate School of Library & Information Science
Center for Informatics Research in Science & Scholarship
University of Illinois at Urbana-Champaign
cathmps@illinois.edu

ABSTRACT
Data-intensive research promises advancements in scientific knowledge and brings with it new demands for staff that can manage large and complex data, design user services, and facilitate open access. Research institutions are extending their services to scientific data management. As more organizations extend their operations to research data, an understanding of how to build data expertise into staff and service models is needed. Understanding the unique expertise required and various data roles is key to providing a well-trained workforce that can support data-intensive science. This study examines how an exemplar data center, the National Center for Atmospheric Research, develops and supports their data expertise. This poster reports preliminary results.

Keywords
Research data expertise, Data services, Data centers, Research libraries

INTRODUCTION
At the same time that research is increasingly data and computation intensive, scientists are faced with growing expectations for open access from funders, publishers, and other stakeholders. Access to research data is seen 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, archiving, data sharing, and access. As electronic data and open access expectations grow, research institutions increasingly require a workforce with data expertise, ranging from data analysis and modeling to curation, discovery, sharing, and reuse. Data science and curation are nascent fields, where the required expertise is not well understood and professional boundaries are still being negotiated between various fields including information science, computer science, and domain sciences.

This study examines how organizations develop their own data expertise and design data services, through a case study of a mature geoscience data center, National Center for Atmospheric Research (NCAR).

RELATED WORK
Current trends in eScience are 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). As a result, various 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 (Lyon, Wright, Corti, Edmunds, & Bennett, 2013; Maatta, 2013; Sierra, 2012), and other sectors have reported a shortage of workers with the right skill set for working with data (Manyika et al., 2011; TEKSystems, 2013). To fulfill workforce needs, educators need an understanding of the different data roles and their preparation requirements (Varvel, Palmer, Chao, & Sacchi, 2011).

In the era of data-intensive science and scholarship, organizations are extending their services to research data. Research and data centers have provided valuable data products and services to their domain user communities for decades. More recently, academic libraries are reporting new data service models (Choudhury, 2008; Dasler, Muñoz, & Nilsen, 2013; Ray, 2013; Steinhart, 2011). As organizations respond to the increasing need for data services, they need an understanding of the staff roles and expertise required to support eScience services and how to build data expertise in their organizations. A misalignment of expertise and services could hinder the success of data management services and eScience as a whole. Understanding the unique expertise required in various data roles is essential to providing a well-trained workforce that can support data-intensive science and scholarship.
RESEARCH QUESTION
This study examines how an exemplar geoscience research center, NCAR, develops an in-house data expertise. The poster will address the data roles, knowledge, and skills necessary to support data services.

CONCEPTUAL FRAME
The concept, knowing in practice, developed by Orlikowski (2002) guides the understanding of expertise and competence needed for data work. Orlikowski (2002) proposed the concept of knowing in practice (hereafter referred to as knowing) as a new approach to understanding organizational competence and expertise. This concept focuses on the doing: “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. Organizational expertise is developed by ongoing practices of a group of employees, which may be distributed across departments or locations. In this study, knowing (Orlikowski, 2002) was applied to the analysis of the expertise and competence required for data work and how expertise develops over time in an organization.

METHODS
Employing a case study approach, this study examines data staffing and expertise at NCAR. NCAR has been managing scientific data and establishing their data expertise for several decades, offering an ideal setting to examine the process of data expertise and service development. The NCAR case study was supplemented with interviews of managers at other data centers to extend, enrich, and further validate the case study results.

This study draws on two primary sources of evidence: 1) semi-structured interviews with staff responsible for data services and hiring data workers; and 2) artifact collection of organizational reports, policies, charts, job advertisements, and related materials. A pilot study was conducted at NCAR in Fall 2014. To date, 11 interviews at NCAR have been completed with supplemental organizational documents collected from the NCAR archive and website. An additional set of 19 semi-structured interviews was conducted with managers at 18 different geoscience data centers.

Qualitative data were analyzed using ATLAS.ti 7. Analytical codes were developed using both an inductive and deductive approach. Data were coded initially to identify emerging themes. Next, the author reviewed the research questions and literature to generate additional codes. A codebook was created with the final set of codes for analysis and their definitions. The author coded the data.

Case Description: NCAR is a premiere 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 to include 8 labs and programs conducting research on topics including sun and earth connection, air quality and chemistry, and climate systems and interactions. NCAR is a federally-funded research and development center, funded by the National Science Foundation, supporting research and providing instruments, field campaign support, supercomputing facilities, and archival collections of data for the atmospheric science community (Mayernik et al., 2014). The center 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), and Community Data Portal, as well as individual lab and project websites, real-time data feeds, and external domain repositories.

PRELIMINARY RESULTS
The results are reported in three main areas: participation demographics, data expertise, and data staffing and roles.

Participant Demographics: At NCAR (N=11), the interviewees held a variety of positions including Project Scientist, Database Engineer, Software Engineer, and Manager. The average length of time in the position was 8.6 years (min=2, max=24). All participants were working in field of the atmospheric and climate sciences.

At other data centers (N=19), the participants held positions with a variety of titles such as Research Scientist, Information Specialist, Information Architect, and Data Manager. The average years employed in their current position was 9.9 years with a range of 2 months to 20 years. The domain of work varied: Ecology (5), Oceanography (3), Geology (2), and Urban Science (2). The following sub-disciplines each had one participant: space, environmental, geophysical, glaciology, computational/informatics, and interdisciplinary.

Data expertise
The informants were asked to describe the data expertise and the unique contributions of data professionals within their institutions. Common themes, mentioned by all interviewees, were managing scientific information, understanding the end users, harnessing technology, and standardizing metadata and data work.

NCAR has developed an understanding of how to manage scientific information, data, and other research products. Key information management activities include: metadata, quality assessments, preservation, search and retrieval, data structures (e.g., XML, NetCDF, databases), and data manipulation or transformations. Data professionals have the unique ability to organize these data and products into valuable, useful collections or systems for communities. One participant compares data professionals to librarians:
“In some ways, we are librarians for large archives of information. We try to manage and make that information accessible just like you would in a library with books” (ID#13). All the interviewees, coming from geoscientific backgrounds, described learning how to manage data by doing it. For instance, a data engineer described how s/he learned to work with structured data on the job and primarily by trial and error, where s/he received data in different formats and had to prepare it for deposit into the archival database. NCAR has cultivated a staff with deep knowledge in how to work with different data structures and domain knowledge of common data types and issues, preparing them to perform data management in the atmospheric sciences.

Understanding the user is a critical area of expertise for data sharing and reuse that NCAR has developed in their staff. Data professionals recognize the potential uses of data sets and the needs of end users, making the connection between use and mode of access. One manager described the ability as “putting yourself in the scientists’ shoes to recognize what’s needed in terms of service” (ID#11). Several interviewees mentioned blending the knowledge of the domain with the ability to design user-friendly data portals and applications for data access, use, and analysis. For instance, a data manager described how she came to her position with the domain knowledge but has learned the website and database design on the job through “…trial and error. Often other staff cannot answer my questions, and I spend a lot of time going around asking people” (ID #1).

This quote illustrates the social and experimental dimensions of learning that has built and supported data expertise at NCAR.

NCAR has become a leader in developing scientific software, structures, and tools such as Earth System Grid and NetCDF format. Data staff members have expertise in harnessing technology for scientific data management solutions in areas of storage, dissemination, and movement. Several participants described how NCAR has staff with expertise in software engineering, enabling them to design new applications when off-the-shelf software does not meet their data needs. NCAR also has cultivated staff with considerable experience in utilizing a variety of software packages and scripting languages (e.g., MatLab, R, Python). A few seasoned 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. These new skills often were learned from other NCAR staff, trial-and-error, or continuing education courses through NCAR. NCAR has developed staff with knowledge and experience in varied technologies and with the ability to design computational solutions.

Standardizing metadata and data is an area where NCAR staff members have been building their knowledge and expertise. The development and use of common standards or best practices is gaining importance in the atmospheric and climate sciences as well as other domains. All participants described how data professionals provide a unique knowledge and experience in the application of metadata and data standards including Federal Geographic Data Committee (FGDC), ISO 19115, Ecological Metadata Language (EML), DataCite, OAI-PMH, and OAIS. One data manager summarized how “…you have to know how to translate from what you’ve got from that laboratory person in the first step and put it into something that some other catalog like NCAR or anybody can use” (ID#126). A few data managers highlighted how serving interdisciplinary communities requires knowledge of multiple standards:

“I see the need for us to be putting out the metadata in various ways so when we’re working with the general public then it might make sense that we’re providing for instance the core elements that you would find in the Dublin Core…when we are trying to get information out to the earth science community, having it in FGDC…if you’re getting it out into other catalogs for example having it in the DataCite format.” (ID#116)

This quote illustrates the importance of audience in metadata generation and relationships between metadata standards. An understanding of metadata quality was another critical area of expertise frequently mentioned by participants. A data scientist described how data professionals understand, “what makes a good title, what makes a good metadata record” (ID#114). The knowledge and compliance with standards and best practices are an expertise that NCAR has nurtured in their data staff.

Data staffing and roles
NCAR contains multiple data management teams or units across its labs. Labs at NCAR are classified by the sub-disciplines of atmospheric science (e.g., climatology, atmospheric chemistry), and each lab has a unique configuration of staffing for data management. For instance, the High Altitude Observatory (HAO) embeds data managers into research teams, while the Computational Information Systems Laboratory (CISL), which houses the Research Data Archive (RDA), is a unit comprised of data experts serving the NCAR community. The NCAR labs may have one or multiple teams providing data services. Across NCAR labs, four main data roles have emerged and evolved over the years: data managers, data engineers, data scientists, and managers.

• Data managers (or curators) are responsible for the curation of research data including tasks such as data manipulation, metadata generation, preservation, dissemination, and user consultations. These professionals are primarily concerned with good management of the data to ensure use and reuse.
• **Data engineers** are building tools and applications for data stewardship such as data capture, discovery, and analysis. The engineers are designing open access software, automating data processes, and developing computation solutions for data providers and end users.

• **Data scientists** primarily build models, simulations, and applications for analysis along with responsibilities for scientific data management. These scientists usually are embedded in research teams and producing valuable, high-level data products for reuse.

• **Managers** provide oversight and direction for data stewardship or archiving units. The managers are responsible for the activities, performance, and outputs of their unit.

The series of interviews with 19 geoscience data centers confirms these 4 data roles. A few data centers have data managers divided into more specialized roles such as metadata, archiving, and processing.

One theme that emerged from the interviews is that data positions are moving from specialist to generalist roles. To keep up with growing demands, data positions need to change because data center managers do not expect funding levels to increase. Several participants discussed how data positions require generalists able to work across all phases of the lifecycle and perform different activities such as data generation, processing, metadata, archiving, and supporting reuse. One manager forecasts the specialist roles will be “eliminated due to either changes in technology or direction” (ID#116). The generalist roles may become more popular as trends and priorities in scientific management evolve.

While each lab has its own data staff and efforts, NCAR has a history of organization-wide initiatives on data management (Baker, Mayernik, Thompson, Nienhouse, Williams, & Worley, 2015). A recent effort is the Data Stewardship and Engineering Team (DSET) aiming to adopt organizational best practices (e.g., DOIs) and an NCAR-wide data discovery system. DSET, started in 2014, is comprised of data representatives from all NCAR labs. The team provides opportunities for data staff to leverage peer knowledge, standards, developments, and systems across labs.

The poster will present a full analysis explicating data service roles and expertise in the NCAR case study.

**CONCLUSION**

eScience promises advancements in scientific knowledge while presenting new demands for data professionals. NCAR has developed an in-house expertise in working with scientific data by embedding knowledge, skills, and roles into their staff and labs. NCAR has cultivated, both at the individual and organizational level, expertise in managing scientific information and data products, understanding user and users, harnessing technology for data solutions, and standardizing metadata and data. These areas of organizational expertise overlap with the trends in data curation workforce needs (Mayernik et al., 2014; Palmer, Thompson, Baker, & Senseney, 2014; Swan & Brown, 2008). NCAR units have evolved over the years to create roles for data managers, data engineers, data scientists, and unit managers, embedding data knowledge and skills into each lab. NCAR also has supported organization-wide efforts for data management, leveraging knowledge, service models, and systems across all the labs. Based on preliminary results, NCAR provides a model for how research institutions can build data expertise into their staff.

Data collection for this study is ongoing. The author anticipates that the results will help answer questions on how organizations can develop data expertise. The case study will inform further development of a framework of data knowledge and skills needed to offer particular data services. The preliminary results have given insight into a set of core areas for data expertise and roles. These results are important for research institutions and educators planning data curriculum.

**ACKNOWLEDGMENTS**

This research was supported by the Data Curation Education in Research Centers (DCERC) program, funded by the Institute of Museum and Library Services grant (RE-02-10-0004-10).

**REFERENCES**


http://www.jisc.ac.uk/uploaded_documents/e-
scienceReportFinal.pdf


