DATA WORK CONFIGURATIONS IN THE FIELD-BASED NATURAL SCIENCES: MESOSCALE INFRASTRUCTURES, PROJECT COLLECTIVES, AND DATA GATEWAYS

BY

KAREN S. BAKER

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library and 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 Michael B. Twidale

ABSTRACT

This multi-case, longitudinal ethnographic study investigates data work configurations of research projects in the field-based natural sciences. Project collective data work involves managing data in addition to facilitating digital data archiving. Through qualitative analysis, the concepts of data work arenas, information environments, and pre-archive data work are incorporated into a Data Work System model that foregrounds mesoscale infrastructures central to the movement of data from its origin in the field to its destination in an archive. The model integrates a system perspective within which data intermediaries play a key role as infrastructure is grown to support the dynamics associated with research data use. As an outcome of the analysis, three kinds of mesoscale data collectives are characterized as Local, Archive, and Developing. Three case studies illustrate the diversity of data work configurations, characterize mesoscale infrastructures as future-making prototypes, and demonstrate the relevance of Local Data Collectives as Data Gateways in planning information architecture. The cases contribute to the development of conceptual resources critical to maintaining the vibrancy and vigor of scientific research and the data work associated with data production in addition to knowledge production.

ACKNOWLEDGEMENT

I would like to thank my committee members – Carole Palmer as chair and Joel Cutcher-Gershenfeld, Matt Mayernik, and Mike Twidale - all of whom made time for discussions whenever needed. Each contributed in significant ways and from different perspectives during my studies and dissertation work.

Participation in the Center for Informatics Research in Science and Scholarship (CIRSS) was a special opportunity to venture into the tangled realm of data practices and data concepts. My work was informed by the NSF DataNet award for the Data Conservancy: A Digital Research and Curation Virtual Organization (OCI-0830976) and was supported by the Data Curation Education in Research Centers (DCERC) project funded by the Institute of Museum and Library Services (RE-02-10-0004-10). DCERC supported my fieldwork and mentorship at NCAR. The School of Information Sciences with its faculty, staff, resources, and approach created an environment of respect for the diversity of perspectives in the information sciences.

I deeply appreciate the kind words and companionship of family, friends, colleagues, mentors, and fellow students during these years of study and reflection on scientific data work and information environments. My information management work while at University of California San Diego was informed by interdisciplinary collaboration within the Long Term Ecological Research (LTER) program, the Ocean Informatics Initiative at Scripps Institution of Oceanography with Jerry Wanetick, the UCSD Science Studies program, and studies of scientific infrastructure with Geof Bowker, Helena Karasti, David Ribes, and Florence Millerand.

I was fortunate with the research cases I investigated. Participants and data allies including Nicole Kaplan, Dan Draper, Matt Mayernik, Steve Williams, Mike Lemke, Rip Sparks, Doug Blodgett, and Jeff Walk were generous with their time and provided access to project workspaces. In addition, their care for both natural and digital ecosystems was uplifting.

Work on this study has been a challenge and an adventure that provided me insights professional and personal. I acknowledge those who came before me with sensitivity to new roles and to the value of diverse teams. I also recognize the good fortune that provided me the opportunity to create this body of work and to continue learning.

TABLE OF CONTENT

EPIGRAPH

A voice of experience on the architecture of buildings provides guidance pertinent to design of information architecture inclusive of a diversity of views and the art of science:

“I speak of a complex and contradictory architecture based on the richness and ambiguity of modern experience, including that experience which is inherent in art.” (Robert Venturi, 1966)

CHAPTER 1. INTRODUCTION

Information infrastructure, an evolving concept in contemporary development efforts, is intended to provide support for scientific research and its data. A study of data work currently carried out by scientific research projects can contribute to our understanding of information infrastructure support of inquiries in the field-based natural sciences. This study focuses on collective, digital data work and project-repository relations in order to characterize the kinds of data work configurations that now exist and to consider the elements of infrastructure supporting them. Data work configurations involve both the relations between scientific work and data work as well as between projects and data repositories. A three-component framework and a data work system model are developed for the earth and environmental sciences to demonstrate work with data as it is moved from the field to local arenas and data repositories. This multi-case, longitudinal ethnography of scientific project data work builds on research in a number of fields that addresses sociotechnical systems, the growth of information infrastructure, and data care. Collective data work of projects as mesoscale undertakings has been explored within the larger context of a data work system.

The current state of support for data is reviewed below, followed by my research questions and research contributions. The final section provides a brief overview of the chapters in this dissertation.

1.1 Supporting Scientific Data in the Digital Age

Dramatic changes in data work since the advent of the digital era have impacted scientific research projects and data repositories. Following the rapid increase in digital capabilities across the United States over the last decades, an executive mandate in 2013 to federal agencies called for open access to data associated with federal grant awards (Holdren, 2013). It resulted in a number of agency policies that targeted requirements for data sharing by researchers. The mandate represents a new social contract for science that has spurred change in data practices. It is a major shift from the traditional view of science as conducted by individuals and laboratories to a vision of a data landscape featuring open data and open science (TRS, 2012; Pasquetto et al., 2016; Kowalczyk and Shankar, 2011). This vision spotlights science as a source not only of knowledge that reaches colleagues through disciplinary journals but also of data destined to

reach diverse communities. For this endeavor, information infrastructure is being developed to support *research* data and work with research data. This infrastructure will be referred to below in short as data infrastructure.

Information infrastructure, a concept that brings together information systems and networks, requires further definition when used in practice. It is seen in a variety of ways as providing, for instance, an information window, a communication tool, and a data pipeline. Designers of information infrastructure plans are grappling with the unprecedented scales, scopes, and dynamics of digital architecture. Infrastructure exists at different scales ranging from small to large, sometimes referred to as micro approaches and macro approaches to information systems. The term mesoscale references a relative scale, fit to purpose for the object and context of interest. In this investigation where field measurements are made by an individual or small group using an instrument described by a digital manual, both instrument and manual represent small-scale information infrastructure. Project-specific data collectives that archive data generated in the field by project researchers are designated mesoscale infrastructure. At the macroscale are archives with extensive geographic reach, thematic breadth, and/or diversity of scope.

In the field-based sciences, data frequently is managed at the level of a research project as scientists work collaboratively to understand the natural environment. A research project is a unit of organization that differs from those at the individual, laboratory, or enterprise level. Research projects are typically inquiry-driven efforts that generate data from observations and measurements of the earth and the environment. The group of project participants conducts joint field campaigns and manages data collectively in various manners. In addition to being a source of data, projects often create an integrative environment within which information abounds and knowledge is created collectively. Project data was assembled and documented as a collection of inter-related datasets and associated project artifacts. The concept of an ‘information environment’ emerged in this investigation of project data work. Taking the assembly and management of data for field science projects as a matter of concern supports ongoing efforts in building digital data systems and information environments and informs the development of digital data infrastructure within the sciences.

Data repositories of many kinds are being developed in response to the volume of data and the number of associated data services required. Transitioning to the ‘Age of Digital Data’ is

revealing complex issues that arise as repositories are developed for research data and interwoven with existing practices. Research responsibilities are in effect expanding to include awareness of and planning for data management and data systems that support assembly, preservation and reuse of data. An ARL report (2009) makes the point, however, that 'repositories are developing rather than developed'. Though the characterization of data repositories is still developing (Baker and Duerr, 2016b), there are three distinct kinds of repositories in this study: 1) project-specific, 2) disciplinary, and 3) institutional. Vocabularies for these data efforts are still growing in order to accommodate concepts such as data access, data production, and data infrastructure.

This study demonstrates that data work configurations and the support they provide vary across the sciences. From a mechanistic or positivistic perspective, digital data is an immutable resource that travels via systems configured to support automated, seamless flows of data. The number of data systems and repositories as well as the scope of data work are increasing, however, making it evident that rapid convergence on 'a solution' to data access is unlikely given the differing goals, methods, standards, and vocabularies together with limited budgets at play. New approaches to supporting scientific data and its management are undergoing development but often lag what is envisioned in information infrastructure plans. A richer understanding of data arrangements emerges when project-related data work is studied as data moves through the multiple work arenas where it is managed in the form of files, sets, collections, and packages with a variety of formats.

1.2 Research Questions

From analysis of interviews, observations, and project materials, I investigated two research questions pertaining to field-based project data. The first question focused on field-to-archive data work configurations:

> RQ1: How do differing configurations of scientific research projects and data repositories support the movement of data from projects to data archives in the field-based natural sciences?

With project data aggregated in multiple ways and moved across many work arenas, related

questions include: “How do the project-repository relations differ?”, and “Why are project-repository configurations different?”

The role of projects in assembling, organizing and managing collective data that ultimately is deposited into an external data archive was also investigated. While the first research question scrutinized data work configurations, the second research question focused on the data work:

> RQ2: What are the characteristics of data work and elements of infrastructure that enable the movement of data from projects to data archives?

By focusing on the description of data work and infrastructure, differences in data arrangements can be highlighted. Related research questions include “How do the characteristics of data work vary?” and “Why do the elements of infrastructure differ?”

1.3 Contributions of this Research

Intellectual contribution

In a data landscape populated by diverse kinds of data and data work, the examination of existing data work configurations is needed to inform the development of a contemporary data infrastructure designed to support both project and preservation goals. This in-depth investigation of data work from the perspective of scientific projects engaged in generation and assembly of data, contributes to the discourses on information infrastructure as well as on data management, data curation, and data repositories. Given the tendency of data to disappear into individual laboratories at the end of project funding, each field-to-archive configuration prevents data diaspora. This account is intended to be a step towards a deeper, empirically grounded, understanding of data work and the growth of data infrastructure.

The mesoscale for a project data work configuration is a transitional point between the data origin where researchers generate data and the larger scope of digital data archives. Few past studies of data practices have explored existing data arrangements of research projects though the concept of ‘the middle’ in science and technology has been explored (Wyatt and Balmer, 2007) and the relational nature of arrangements highlighted as ‘in-between infrastructures’ (Botero and Saad-Sulonen, 2010). My research adds to the body of data

infrastructure studies, expanding upon work that often focuses on lifetimes of data within an archive, removed from the research projects that generate the data. As 'in-between infrastructure', the focus is on data work at the project level rather than the work at more remote data destinations with larger scopes such as institutional repositories and archives or the work of cyberinfrastructure, computational grid, and high performance computing endeavors.

The aim here is not to propose an overarching theory. Rather, the contribution of this work is as a qualitative inquiry into existing data arrangements with a focus on a project's collective data work. This research identified some of the elements of data infrastructure that address the data deluge within local venues. The role of data intermediaries is observed in conjunction with distributions of responsibilities for data work. In addition, various trade-offs are found to shape the infrastructure supporting project data work. Together with others, the Belmont Forum on e-infrastructure (Allison and Gurney, 2014) called explicitly for cases to add to our nascent understanding of data infrastructure. My study adds to a growing set of case studies using ethnographic methods to contribute to our understanding of the work associated with data and information infrastructure.

Practical contribution

With increasing awareness of the potential for technology to become a dominating force, the descriptive approaches of ethnography are an informative strategy as well as a liberating strategy that avoids the twin dangers of macroscale bias and marginalization of local contributions in the study of infrastructure. The in-depth description of cases of data and repository arrangements for field-based scientific projects provides insight into the kinds of data work configurations and local capacities that contribute to the production of data and data sharing efforts. An examination of the characteristics of field-to-archive data work and elements of data infrastructure enables the identification and categorization of mesoscale data collectives that support joint work with project data.

The notion of collaboration via data sharing creates new expectations for scientific research and often underestimates the need for design and management in support of collective data work that occurs in many arenas. The 'data sharing' spotlight illuminates the digital landscape, so that invisible data work, underdeveloped vocabularies, and inadequate conceptual development can be addressed as part of a collective consciousness of managing data. There is a sense of urgency

about working with data today because the data that undergird scientific research are being called upon to support societal policy-making at the scale required for a 'managed earth' before data infrastructure is well characterized, operationalized, and enacted.

1.4 Organization of the Dissertation

This dissertation is organized in seven chapters. Having laid out the research questions in the first chapter, Chapter 2 delineates the scope of this research and provides definitions of terms. Chapter 3 presents research methods including the concept of stories of participation and the limitations of this investigation. I describe the three cases studied in Chapter 4 that provide concrete examples of data work configurations. In Chapter 5, a three-component analytic framework provides the context for exploring the local component in data work configurations. Elements of data infrastructure that support local data collectives are identified through cross-case analysis. In Chapter 6, a system lens is adopted that takes into account the interactions and overlap of framework components. A dynamic data work system model inclusive of many data work arenas is developed. A number of elements of data infrastructure for data collectives are explored as trade-offs. Chapter 7 concludes the dissertation with a summary of findings, a review of the research questions, observations on the changing data culture, and comments on potential future directions of this research. Appendices contain interview instruments for the Institutional Review Board (Appendix A), timelines and extended documentation of project data work for three cases (Appendices B-D) and an external archive (Appendix E), information on the LTER Information Management Committee history and guides (Appendix F), and data management position advertisements (Appendix G).

CHAPTER 2. BACKGROUND

Scientific researchers are faced with new concepts and terms as they plan projects not only for data collecting and analysis but also for archiving data. Some of the language associated with grouping data files, assembling data, and carrying out data activities is presented below prior to considering the concept of data work and data work arenas. Following this, scientific research projects are discussed as information environments, information communities, and sites of collection making. The chapter ends with a discussion of the role of infrastructure and design in collective data work.

2.1 The Language of Data

There are a number of key terms and concepts associated with data work of scientific projects. Data management planning requires scientific researchers to put into words how they conceive of their data as well as how they plan to manage and share datasets despite the lack of established concepts and shared vocabulary to draw upon. With the aim of making the handling of data for reuse more explicit, language is being developed through ongoing research where fundamental concepts relating to data and data repositories are the subject of investigation. Such terms are deceivingly simple, requiring great care to explicate effectively for situations involving data from more than a limited number of researchers.

2.1.1 Data and data related terms: Datasets, data packages, and data collections

With concern about data and data-related activities increasing, terms such as datasets, data packages, and data collections are used both informally in local groups and are defined more formally in various venues. For this study, the terms were used to refer to that which participants consider them to be.

Data is defined in many ways (Borgman, 2015). The Consultative Committee for Space Data Systems (CCSDS, 2012) provides an operational definition of data: “A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen.” In the

field-based sciences, data includes empirical observations and measurements often described as factual and used to support research findings. The term is used in the singular to refer to a concept or a group of items and also in the plural when emphasis is on the individual items in a group.

Dataset is one of the terms used to describe a grouping of data. It was found in the literature to have multiple meanings (Renear et al., 2010) that involve grouping by content, purpose, and relatedness. A few definitions were identified in the cases studied. For example, dataset might be used to describe a group of data files taken during the same field visit, a number of similarly structured data files, or a group of variables present in a single data table.

Data packages serve as a mechanism for standardizing data delivery and data system ingestion. A 'data package' is similar to the information package defined by the Open Archive Information System (CCSDS, 2012). It contains the recorded digital data sent from a data source to a repository. In general, a package in this study included two kinds of items: first, one or more data files and second, a set of metadata materials in a metadata folder. Zipped into a single file, a package contains the data files and at least some minimum descriptive information necessary for a basic understanding of the data files. The metadata folder includes various descriptive materials such as a README file, a metadata file, a variable definition file, and related supplementary files such as photographs.

Definitions for data collection, another term used when grouping data, are twofold. It is used to describe an activity as well as an entity. As a verb describing action, it refers to collecting or generating data during fieldwork. In this study it is primarily used as a noun that refers to a set of data objects related to a project, though it was also used to refer to other themed collections designating where all the data has been generated such as a field visit, a particular instrument, or the output from a defined analysis process. Within library and archive sciences, the term 'collection' historically refers to aggregated holdings often grouped on the same shelf or in the same box. For this study, a project collection is defined as a purposeful assembly of project datasets and project-related materials. When developed by project participants, it is a primary source.

2.1.2 Data systems, repositories, and archives

A collective effort to assemble and provide access to data involving digital technical

systems is described by terms such as data system, data repository, and data archive. The use of these terms varies across sectors, disciplines, and organizational units in the sciences. A data system at minimum refers to one or more applications that store and manage data as a collection. Repository is generally a broader term that refers to a cache of data with associated services that support registration, storage, management, and documentation of digital materials. Registration involves the assignment of local or global identifiers that establish the unique identity of ingested data. Data system, repository, and archive functionalities include discovery, search, query, and delivery in addition to supporting network access to digital materials. In addition, automated online services may be provided such as a variety of upload capabilities and metadata validation services.

Digital repositories have been described based on a number of characteristics such as their organizational type (e.g. Schmidt, 2010) and their size (Pomerantz, 2008). An early survey proposed two categories of computer-usable archives: 'general purpose service' and 'local service with limited access' (Bisco, 1967). While Cragin et al (2010) explored institutional repository contributions to research (2010), Armbruster and Romary (2010) suggested four categories of repositories: national, institutional, subject-based, and research. Baker and Duerr (2016b) report an even greater variety of kinds of data repositories. Efforts to categorize repositories will be informed by the more inclusive approach of an international research data repository registry, an effort launched in 2012 that supports self-registering of digital data repositories (re3data, nd; Pampel and Dallmeier-Tiessen, 2014). With 1500 repositories registered in 2017, the registry is using this broad base of information to develop a vocabulary of required and optional properties to describe a repository's "general scope, content and infrastructure as well as its compliance with technical, quality and metadata standards" (Rücknagel, et al., 2015).

An archive is generally considered a particular kind of repository with highly structured services that may include formal certification. The mission of an archive explicitly includes long-term preservation. As with data systems and repositories, there is variation in the definition and categorization of archives. The Reference Model for an Open Archival Information System (OAIS) defines an archive as "an organization that intends to preserve information for access and use by a designated community" (CCSDS, 2012). The online Consortia Advancing Standards in Research Administration Information dictionary of terms (CASRAI, nd) defines an archive as "a

place or collection containing static records, documents, or other materials for long-term preservation."

2.1.3 Data curation, data stewardship, and data management

Use of the terms data curation, data stewardship, and data management overlap in ill-defined ways. A number of models and definitions have been used to convey and characterize the concept of data curation. Lord and MacDonald (2003) provide a series of models illustrating the development of data curation roles, activities, and products over time. Their definition involves different levels of data curation: "The activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and with other published materials." Data curation is defined succinctly in the library realm as involving "ways of organizing, displaying, and repurposing preserved data." (ARL, 2006). Alternatively, it is defined as "the active and on-going management of data through its life cycle of interest and usefulness to scholarship, science, and education" (Palmer et al., 2008). Palmer, et al. (2013b) present a history of data curation as both a concept and a field with a view of data and information curation as "purposeful work". It is the purposefulness of this work that distinguishes it from more general collecting activities (Palmer, 2004).

The concept of stewardship has a history prior to its use in referring to concerns for care of data. Stewardship appears in biblical references relating to responsibility for managing the earth, in library references relating to preservation of documents, and in environmental calls for responsible use and protection of the earth. The National Academies of Science report (NAS, 2009) is one of many reports on the utility and integrity of research data does not use the term stewardship but addresses many aspects of stewardship including selection of data, discovery of errors, breaches of trust, and issues of credit. Stewardship of digital data, however, has emerged from discussions of the challenges associated with long-term data as a term that refers to an ethic of planning and providing active care for data that reaches beyond sole ownership to data resources as a public good (Duerr et al., 2004; Karasti et al., 2006; Baker and Yarmey, 2009; Palmer et al., 2013b; Hartter et al., 2013; ICPSR, nd).

As attention to data curation and data stewardship has increased in recent years, so has attention to data management (e.g. Pryor, 2012; Strasser et al., 2012) and the 'data lifecycle' (Higgins, 2012; Humphrey, 2006). Different communities conceptualize these terms in different ways. Data management is defined as "an active process by which digital resources remain discoverable, accessible and intelligible over the longer term, a process that invests data and datasets with the potential to accrue value as assets enjoying far wider use than their creators may have anticipated" by Pryor (2012). This definition is close to many of the definitions for data curation with its added proviso of a long-term period for preservation and dissemination. When the work with data includes the planning, structuring, organizing, and reporting that facilitates the interpretation of data, it is sometimes referred to as information management. Understandings of and familiarity with these titles differ. One data manager (DM) in this study working with a field-based project while involved in assembling data for immediate use by project researchers as well as for overseeing the submission of data to an archive, explained how the use of titles relating to these terms changed in practice depending upon the audience first in order to convey the work at hand to those familiar with its generation, interpretation, and use in the field and then to those familiar with its registration, preservation, and reuse at an archive:

> I'm the information manager. It's funny because I keep going back and forth between information manager and data manager. Because you know when I speak to scientists I call myself a data manager. When I speak to the digital curation people, I call myself an information manager. (DM)

After 20 years of expanding project-related responsibilities, the role once recognized within the Long Term Ecological Research (LTER) Program as 'data management' was changed in 2000 at the suggestion of the network-wide Data Management Committee to 'information management' (Baker et al., 2000). This committee of information managers embedded in their separate project sites was an active community of practice. A brief history of the committee and an overview of the information management guides they developed are given in Appendix F.

2.1.4 Data work, data products, and data work arenas

The concepts of data work and data work arenas are introduced to open up data-related discussions. Data work is defined here as any effort applied to data including when it is created, recorded, managed, modified, used, transformed, preserved, modeled and/or made accessible.

These diverse activities are carried out in many settings and involve a variety of tools. The term is a broad category for the many activities sometimes referred to as data management, data curation, or information resource management (e.g. Van Den Hoven, 1999; Cragin et al., 2007; Palmer et al., 2013a). Whether the intent is explorative inquiry or a planned task, a data work outcome is called a data product that manifests in a variety of forms including a table of numbers, a visual representation such as an image, or a set of codes such as an index used to express a summary of a group of data. Note, however, that data work, with its focus on data, is considered distinct though related to technology-driven work associated with data systems. The notion of data work complements that of data practices, a term referring to data-related actions that take place within existing conventions and rules. Data work highlights the labor and responsibilities associated with data practices including the effort involved in bridging human-nonhuman interfaces.

In conjunction with data work, the concept of a 'data work arena' is introduced to characterize a setting where groups of data work tasks representing key process areas are performed, often with a distinct vocabulary developed to describe information relating to the tasks including their phases, revisions, refinements, troubles, and status. Some data work arenas may be confined by where and when the processing occurs such as that carried out in the field versus postfield activities in an instrument lab. They may also be bounded in other ways such as by work with particular kinds of datasets or by association with a particular processing approach. The individuals who carry out the work and those with whom the work and outcomes are discussed, share familiarity with particular data work arenas as a group. This small community is generally smaller in size and scope than the social units described as social worlds (Strauss, 1978), social arenas and worlds (Clarke, 1991), and thought worlds (Dougherty, 1987). Recognizing the diversity of worlds configured in ways contingent on spaces of many kinds (e.g. Lindley et al., 2017), ensures a continuing ability to respond to the heterogeneity of data and data work in field-based research.

2.2 Scientific Research Projects

Scientific research projects have been recognized as a traditional unit of organization in the field-based sciences. The National Science and Technology Council (NSTC, 2009), an early

report describing a strategy for preservation and access to data online, noted projects as entities having a role in harnessing data alongside efforts of the government as well as organizations and domains. The report is silent, however, in defining that role or the work associated with research data. The concepts introduced below – information environments, common information spaces, and information communities – capture some of the dynamics of data work. These concepts exist within a larger context of ongoing events including changes in organizational structure (e.g. library services, organizational arrangements, research missions, and agency programs), community membership (e.g. shifting department members, staff, partners, and networks), technology resources (e.g. field instruments, hardware, software, and systems), as well as funding agency support (e.g. NSF, NASA, ARS).

Project, a general term used to describe individual and multi-investigator undertakings with a defined purpose, refers in this study to a collaborative research unit involved in assembling data collectively for the purpose of supporting project-related knowledge and data work. These projects traditionally exhibit some degree of coherence in their project vision and data practices that includes joint understandings of data sampling, data practices, and data use. Given the 2013 requirements for data sharing, project participants must now also reach a common understanding of data assembly and archiving. The idea of a project as a community of practice that creates an information environment as a common information space is developed in the sections that follow.

For the cases investigated, data were generated during joint fieldwork that took place over long periods of time. The field campaigns were of two types: a) single site with repeated sampling at a single, well-defined geographic location and b) multiple sites with sampling across selected locales with targeted features of interest, such as a topographic anomaly or a disturbance like a fire. Project fieldwork was platform-based where platforms included support for logistics and sampling, and often for instrumentation and transportation. The environmental cases were site-based platforms in that sampling occurred at one geographic location with nearby field stations and a project office. In addition, the collective data work occurred locally. A site particularly suited to ecological research has been referred to as place-based (Kingsland, 2010). Recently, the concept of site-based data curation has been presented in referring to data curation that focuses on a particular scientifically significant site (Palmer et al., 2017). The atmospheric

science case studied, in contrast, involved use of aircraft that allowed project participants to reach many different sampling locations.

2.2.1 Projects as information environments

When digital support was initially recognized as technology-driven or content-driven, interest in and the value of supporting 'the user' developed over time. The concept of an information *use* environment (IUE) inclusive of system designers *and* users was developed to direct focus to user information needs and information use (MacMullin and Taylor, 1984; Tayor, 1991). The Computer Supported Cooperative Work (CSCW) community also took an alternative approach by highlighting the cooperative work supported by technology (Schmidt, 2010; 2011). Bannon and Schmidt (1991) proposed the 'shared information space' as a concept that expands upon traditional group communications to include shared concepts and expectations of participants. This developed into the 'common information space' described by Schmidt and Bannon (1992) where "The work involved in both putting information in common, and in interpreting it, has often not been sufficiently recognized." Collective data work with field-based projects fits within a collaborative construct such as 'common information space' (CIS). This space provides a place for negotiation of interpretations where differences are identified and sometimes resolved and sometimes sidestepped. A CIS can enable the sharing of the meaning of objects and facilitate cooperative decision-making. Bossen (2002) elaborates on the CIS while Rolland et al. (2006) found that large-scale CISs tended to reproduce fragmentation as an unintentional consequence of integrating heterogeneous sources of information.

Scientific research projects may be considered communities of practice that create information environments that are a combination of resources and technologies, participants and practices, procedures and data arrangements, as well as shared experiences and knowledge. A data-related information environment represents a communication arena that brings together shared concepts and expectations as well as an assembly of project data and information. It enables sharing the meaning of objects and facilitates cooperative decision-making among participants familiar with the origin, context, and politics of project data and knowledge. Collaborative project activities that contribute to a project-related information space include shared events and artifacts as well as interpretive and translational efforts. In negotiating accounts attributed to the project, a more nuanced understanding of various views is developed.

The co-construction of meaning by project participants takes account of various views and may result in a hybrid view. Differences that are identified may be resolved or perhaps collectively sidestepped for agreed upon reasons. As a project-related space, an information environment supports knowledge generation (Baker, 2005). Community-aligning activities include mutual decision-making critiques, perspective awareness, and adequate contextualization. The need to recognize and plan for articulation work – the work of scheduling, communicating, organizing, maintaining, fixing, etc. – is generally recognized as playing a significant role in information environments and spaces alike (Bannon and Bodker, 1997; Baker and Millerand, 2007; Boden et al., 2014).

2.2.2 Projects as information communities

A variety of kinds of communities have been described over the past decades. Communities of practice, of interest, of inquiry, of discourse, and of learning have been identified (Fischer, 2013; Carroll, 2007; Lave and Wenger, 1991). In an approach that highlights information and its use, Fisher and Durrance (2003) developed the concept of information communities that exist in diverse settings and are created by collectives. Five characteristics of information communities were identified: 1) Exploit the information sharing qualities of technology; 2) Emphasize collaboration among diverse groups that provide information; 3) Anticipate and often form around people's needs to access and use information; 4) Remove barriers to information sharing and community participation; and 5) Foster social connectedness within the larger community. Rather than focusing on an information environment, Fisher and Durrance (2003) focus on the community associated with an information environment explaining that: "The interaction between the information needs of people and the available content provided via the information system described is the genesis of an information community."

Communities are sources of innovation. Schumpeter (1939) (from NSTC, 2009) defines 'innovation' as the creative use, modification, or combination of existing concepts and devices for desired applications." The National Science and Technology Council report (NSTC, 2009) underscores the value of communities for discovery and innovation: "This diversity in data, individuals, institutions, disciplines, contexts, and cultures is a strength of the American scientific research and education system. One-size-fits-all solutions must be avoided. Solutions should support communities of practice and leverage their capabilities while promoting data

integration and interoperability."

A scientific research project in sharing goals, resources, and information becomes a community that often contributes to and supports an information environment. Like work groups described as multilevel systems (Kozlowski and Ilgar, 2006), a project has members with complementary skills that support participants as they perform activities and share responsibilities. In addition, project participants in the natural sciences frequently share field experiences and are familiar with the various circumstances of data generation. For the long-term, project-based cases in this study, change and innovation are stimulated by the joint planning and conduct of fieldwork as well as associated data work that often involves adapting to new insights, approaches, technologies, and resources.

2.3 Collections and Collection Making

The definition and use of the term 'collection' sets the stage for the notion of project collection making. A collection is a group of records, an aggregation of physical or digital objects gathered together purposefully (Yeo, 2012). Collections are recognized as having membership criteria and creating information context (Lee 2000; Palmer et al., 2010) as well as establishing curatorial responsibility (Buckland, 1999). Collections have been described by Wickett et al. (2013) as satisfying three constraints: 1) collections have member items that have been aggregated; 2) membership is determined by some criteria, and 3) a collection may be treated as an individual object. Further, membership of items in a collection is reported as distinguished by four criteria: context, measure, witness, and aboutness. Items in a collection provide evidence of what is known and of interest about the collection. A National Science Board report (NSB, 2005) discusses functional categories of data collections as well as data authors and data managers but does not elaborate on collection making.

A project's website during this study is a pre-archive, collection-building activity. Website architecture involves assembly and delivery mechanisms that contribute to collection making and identity building by projects. As a result, a collection of artifacts and documents may be held and made accessible by those actively engaged in the process of data generation. A project website serves as a local organizing tool that makes data and related artifacts visible as part of a research project's information environment.

The presentation on the web of information about research projects involves a number of kinds of engagement that contribute to participants' understanding of their work individually and collectively through acts of representation, discernment, and complication (Weick, 2016). For instance, in building their capacity to comprehend meaning and tell a story collectively, researchers interact by discussing and assessing while sorting, filtering, and agreeing on priorities that contribute to development of a community perspective. The process includes negotiating agreement on what is to be included and maintained with the support of data specialists. In speaking of humanities scholarship, Palmer (2004) provides an apt reminder of the scholarly process involved in building (or growing) a collection of digital resources: "... research takes place in the production of the resource, and research is advanced as a result of it." The dynamics of collection making and the visibility of the collection differ from availability of the final collection snapshot at the end of a project. For one project in this study, the assembly of content was subject to the continuing processes of update and assessment so supported collaboration within the project as well as communication with public audiences.

The features of thematic research collections are pertinent to discussions of scientific project collection making as is the concept of 'contextual mass' said to "prioritize the values and work practices of scholarly communities rather than collection size" (Palmer, 2004). Contextual mass describes the density and richness of a collection as "a system of interrelated sources where different types of materials and different subjects work together to support deep and multifaceted inquiry in an area of research" and where materials are selected in a "deliberate" manner. Project websites that assemble a suite of interrelated source materials including field data are approaching the same kind of density and richness as described for humanities collections that are the primary scholarly product. Together with scholarly publications, the incorporation of project-related artifacts into a project collection such as photographs, field notebooks, field reports, changes in sampling strategies, and news about expected and unexpected findings in synthesizing the data can inform current and future interpretation of the data.

2.4 Infrastructure

Infrastructure, a term identified as an area of study (Edwards et al., 2009; Bowker et al., 2010), plays a key role when considering collective data work, data repositories, and project

information environments. As seen in studies of data practices, support for data work varies in environmental science communities (e.g. Borgman, 2012; Cragin, 2009; Zimmerman, 2007; Birnholtz and Bietz, 2003), where both existing and envisioned infrastructure influence data activities (Bowker et al., 2010; Ribes and Lee, 2010; Shaon et al., 2012; Whyte, 2012). Infrastructure research has developed alongside studies of cooperation (Schmidt and Bannon, 1992; Schmidt, 2011a, b), articulation work (Star and Strauss, 1999), and situated work (Suchman, 1987) that has contributed to making visible everyday data practices (Star, 1999). The concept of infrastructure can be recognized for its flexibility (Edwards et al., 2007; Star and Ruhleder 1994, Monteiro et al., 2013; Bowker et al., 2010). It refers to form and content of services in addition to the structure and its support as substrate and suprastructure. Science and Technology Studies and Infrastructure Studies address collaboration at differing levels of scientific organization (Edwards et al., 2009).

2.4.1 Conceptualizing infrastructure

The concept of infrastructure has stretched from an early focus on technical capabilities and arrangements to address concerns with managing data, information, and knowledge (Edwards et al., 2003, 2007, 2013). It continues to develop as our understanding of the complexity of working with communities and digital environments deepens. Studies of telephone and electrical grid have been followed by investigations of digital networks of information systems and help services. In the case for 'how to infrastructure', Star and Bowker, (2002) argue for the need to develop a multi-faceted understanding of infrastructure.

A variety of terms are used in referring to infrastructure, and there are a variety of ways of describing it. The terms cyberinfrastructure and e-infrastructure refer to larger-scale arrangements with computationally and research-based efforts, respectively. Specific categories of systems include those referring to large technological systems (LTS), high performance computing (HPC), and data intensive analysis (DIA). More generally, the human and digital technology components that together form a support base for digital work are referred to as information infrastructure (Hanseth et al., 1996; Edwards et al., 2007). Monteiro et al. (2013) more recently distinguished design for a single location and design that extends to multiple contexts. Schmidt and Bansler (2016) also formulate an alternative approach with three components identified for elucidating kinds of infrastructure: a) computational artifact assembly,

technically situated; b) eInfrastructure, large-scale technical; and c) eInfrastructure, community. In contrast, local collective data work emerged in this study as a dual focus in terms of infrastructure: first, referring to data work and support within a set of arenas and then referring to arenas external to the local component.

Infrastructure provides a way of considering both situated data work arrangements and the infrastructural segments together as a larger infrastructural entity. The concept provides a mechanism for focusing on data work arenas with differing problem-solving capabilities. It makes evident the many boundaries between data work arenas that reside within the same or differing organizational units. Seamfulness is an apt term that suits the work involved in knitting together heterogeneous segments of infrastructures (Vertesi, 2014). The mix of infrastructural segments at different scales, scope, and development phases means the coordination of the segments adds to the difficulty of infrastructure development.

Data work arenas may be characterized by infrastructure dimensions ranging from social to technical and from local to general (Bowker et al., 2010; Monteiro et al., 2013). As the visibility of research data processes and data repositories grows, infrastructure is often targeted as a large-scale construction project. In contrast to pervasive visions of large-scale cyberinfrastructure, there are some descriptions of the social and technical specifics at the local scale from a single project perspective (e.g. Star and Ruhleder, 1996; Benson et al., 2006; Michener et al., 2011; Steinhart et al., 2012; Koch and Chan, 2013; Singh et al., 2012). Recent work highlights social aspects of infrastructure (Wilbanks, 2011; Ribes and Lee, 2010; Bowker et al., 2010; Bowker et al., 1997) that may feature a community or a 'human-in-the-loop' (Mayernik et al., 2013).

The infrastructure for data work as a whole may be described as comprised of nested and overlapping infrastructure segments. Relations abound in connecting up data work in local collective arenas to larger scales. From an Infrastructure Studies point of view, Star (2000) observes, 'it's infrastructure all the way down" drawing on the hierarchical relational regress noted for cosmology by Stephen Hawking who reported, " it's turtles all the way down" (Sandvig, 2010). In considering data movement from field to archive, two ends of a spectrum are identified though my empirical focus is on data work that occurs betwixt and between, at additional stations along the way each of which is a data work arena. These are 'way stations', that is, intermediate assembly locations where data comes to rest along its path to an archive. At

each way station there is an ‘assembly of technical artifacts’ (Schmidt and Bansler, 2016) as well as an ‘ensemble of social practices” (Star and Ruhleder, 1996).

The distribution of scientific data work in multiple arenas is described in many ways. Data work has been described as located ‘upstream’ or ‘downstream’ in reference to the view from close to the data origin or close to the destination of data, respectively (Wallis et al., 2008). In the early or initial stages of data work, a participant at the data origin looks ‘downstream’ and sees a repository in the distance as a data destination. From a destination archive, a participant looks ‘upstream’ and sees earlier stages of the data. Wallis et al. (2008) discuss the need to move archival practices upstream for some kinds of data, while Monteiro et al. (2013) describe ‘local practices’ and Beagrie et al. (2008) create a multi-part activity model that includes a ‘pre-archive’ stage that “primarily relates to research projects in universities creating research data for later transfer to a data archive.” While remote data repositories often work together with communities at the scale of a domain or a country (Treloar et al., 2012; Berman, 2014), a local data repository typically is associated with a specific site or project and serves as an assembly and staging location close to the source of data generation (Kansa et al., 2014; Karasti et al., 2006). Distinguishing local and remote or general as categories for repositories highlights the importance of the data repository position in terms of ‘distance-from-origin’ of the data (Baker and Yarmey, 2009). The local community has first-hand knowledge of and experience with the fieldwork of a single project while remote archives such as institutional repositories and national archives have experience with many projects.

2.4.2 Considering design

Infrastructure development when referred to as ‘infrastructure growth’ evokes an organic unfolding (Nardi and O'Day 1999; Jackson et al., 2007) operationalized by the use of an iterative design approach (Edwards et al., 2007, 2009). A number of design approaches are used in the digital realm today. In the computer and information sciences, spiral design has opened up into a variety of iterative design approaches, including a participatory design focus on continuing design. Regenerative design captures the iterative nature of the work carried out in a variety of arenas that carry forward experience gained from successes and recoveries (Cole et al., 2013; Lyle, 1996). An iterative design approach for infrastructure highlights incremental change as an

alternative to more traditional technical approaches that plan and build infrastructure using a waterfall model.

The active, evolving nature of infrastructure is discussed in the work of Star and Ruhleder (1996) as they studied the complexity of infrastructure and conveyed the notion of infrastructure as an ongoing process involving design, development, maintenance, and update. Infrastructure has been cast into an active form first as a verb 'to infrastructure' and then a gerund form 'infrastructuring' that aims to capture both the work involved in establishing infrastructure but also the ongoing work required to maintain it. While Star reports infrastructure does not develop globally all at once but takes time and negotiation, Suchman (2002a, b) speaks of 'located accountabilities of technical production' and 'artful integration' of collections of technology that lead to local tailoring (Henderson and Kyng, 1991), patchwork prototyping (Twidale and Floyd, 2008) and local enactment (Millerand and Baker, 2010; Baker and Millerand, 2007). Star and Bowker (2002) discuss 'how to infrastructure' while the term 'infrastructuring' is devised for use in the participatory design community (Karasti and Syrjanen, 2004; Karasti and Baker, 2004). More recently the term, 'infrastructuring' is used to capture the ongoing design, maintenance, and redesign of infrastructure over long-term timeframes. Karasti (2014) elaborates on its use while Pipek and Wulf (2009) speak of 'infrastructuring' for the information systems community where they report that infrastructuring evokes "a sociotechnical perspective on the iterative, ongoing processes of designing and using of technological supports for research and its data".

Also related to design, is the concept of alignment that ties structures and practices within a collective of data generators together with those at remote repositories associated with the preservation and reuse of data. Drift resulting in misalignment of data work arenas is a constant concern given change as an ever present factor involving personnel, needs, technologies, systems, and more. An iterative design approach establishes and then aims to maintain the alignment needed both internal to data work arenas as well as between various arenas that enable the movement of data. Alignment represents the 'when' of infrastructure described by Star and Ruhleder (1996). The alignment of segments that connect up with one another, or fail to connect up, occurs in planned and unanticipated ways that require ongoing attention.

CHAPTER 3. METHODS

Methods are presented in this chapter. The first two sections provide information about the research design and research methods and about the study design. This followed by sections on data collection, analysis, and coparticipation. Finally the strengths, limitations, and robustness of the study are discussed followed by a brief summary that includes an identification key for the role categories used to identify participants quoted in the text.

3.1 Research Design and Research Methods

3.1.1 Research design

I take a qualitative approach for this investigation that is framed by the principles of participatory design, ethnography, and case study methods. Qualitative research provides an approach to studying phenomena in their natural settings with the aim of understanding and describing how participants experience and understand the world (Denzin and Lincoln, 1995; Myers, 1999). The case study method provided a way into the layers of data-related activities, thereby escaping generalized assumptions about data work and repository relations.

3.1.2 Research methods

The research methods described in this section include a multi-case approach, participatory design, and ethnographic methods. Ethnographic methods used are described below including grounded theory, situational analysis, and longitudinal ethnography. The section ends with a discussion of constructing the field.

Multi-case approach

Case studies provide an in-depth, holistic view through use of diverse methods and multiple sources of information. A case study is a type of empirical inquiry allowing for what Becker (1970) called ‘rich data’, Geertz (1973) referred to as ‘thick description’, and Latzko-Toth et al., 2017) presented as ‘thick data’ in contrast to ‘big data’. The goal with a case study is to identify current circumstances, understandings, and practices for a close, situated view of the activities associated with a real-life situation where boundaries may not be distinct between the

phenomena studied and its context (Stake, 1995; Creswell, 2007). When carried out longitudinally, a case presents a comprehensive view of a complex system developing over time. A multi-case study asks the same research questions in a set of independent cases. It offers an exposure to a broader range of circumstances, issues, and perspectives. With multiple cases, cycles of cross-case analysis are possible that draw out deeper understandings (Hine, 2007).

Participatory design

Participatory design, a qualitative approach that actively involves stakeholders in the design process, informs investigations of work relating to technologies and their configurations (Van House, 2004). Participatory research in general is an orientation within qualitative studies that involves study participants partnering with a participatory researcher in the knowledge production process and draws on democratic theory, participation, and ethics (Bergold and Thomas, 2012). Qualitative research and participatory research together capture participant perspectives and practices. The approach initially developed in response to issues associated with technology and the quality of life for participants in the workplace (Ehn, 1989; Floyd et al., 1989; Greenbaum and Kyng, 1991). An early cyclic model of Lewin (1946) developed into an action-reflection spiral process with four phases: plan, act, observe, and reflect (McNiff, 1988). Both action research (Reason and Bradbury, 2008) and participatory design (Schuler and Namioka, 1993; Simonsen and Robertson, 2013) involve an iterative or continuing process of interaction. Approaches vary in their focus from empowerment to self-study. While action research aims to prompt change through action and leads to learning-by-doing (Orlikowski 2002), participatory design aims to prompt awareness and understanding that informs participants and leads to action.

Participatory design continues to evolve as it addresses design issues associated with information systems, information environments, and information infrastructures. Studies concerned with activities in the digital realm have changed over time as reported in the work of Bannon (1991) where ‘From Human Factors to Human Actors’ captures a shift in perception of individuals from passive user to active participant. Suchman (2001, 2002a, b), in studying designer-user disconnects, brought forward a view of ‘situated’ work and a model of knowledge that recognizes complex networks of relations between designers, users, and technologies. As an approach to studying situated work, participatory design is a process that fosters critical

reflection on the implications of research results for participants' own work (Schuler and Namioka, 1993). It focuses on collaborative research activities and understandings between researchers and participants who typically in this approach are technology developers, technology users or managers in various settings. Participatory design is described as a methodology foregrounding ethics and sensitivity to dominance in the design of sociotechnical systems (Spinuzzi, 2005). Blomberg and Karasti (2013a) summarize four principles of participatory design as follows: respect for differing knowledge bases, opportunities for mutual learning, joint negotiation of project goals, as well as negotiation of tools and processes to facilitate design.

Concepts such as 'intervention' and 'participation' continue to unfold as they are explored in depth at multiple scales (e.g. Blomberg and Karasti, 2013b; Halskov and Hansen, 2015; Simonsen and Robertson, 2013; Spinuzzi, 2013). Jensen (2007) argues that 'no scale can be the final one' and emphasizes the importance of ongoing collective learning processes. Interventions, the introduction of actions or artifacts that hold the potential to stimulate change, vary in terms of scope and degree. The term 'intervention' describes interactions that range from high levels such as policy (e.g. Jasanoff, 2004) to situated studies in local venues (e.g. Suchman, 2002a, b). Participation, as a form of intervention, includes approaches described sometimes as "modest", "prescriptive", "intervention-as-performance", "translocal" and "co-realized" (Jensen 2012; Heath, 2007; Blomberg and Karasti, 2013a). The participation of an ethnographer may be influenced by factors such as the state or stage of development of a project as well as the circumstances that motivate the engagement of an ethnographer (Ribes and Baker, 2007). Jensen (2012) describes a partnering with participants as 'intervention by invitation' while Gjefsen and Fisher (2014) report that embedding social science into research efforts promotes reflexivity and informs action.

Ethnography

Ethnography is a research method based on fieldwork that allows a researcher to get close to participant experiences and discern patterns in a particular group that shares a project and/or a culture. Ethnography, with roots in anthropology and ethnomethodology, foregrounds the sociality and materiality central to practice theory. Ethnographic methods are appropriate for issues and problems that are complex, beyond our understanding of the dynamics and

interdependencies. Blomberg and Karasti (2013a) describe an analytic ethnographic approach that contrasts with a purely descriptive or scenic report of a community. They state: "First and foremost ethnography is concerned with providing an analytic account of events and activities as they occur, without attempting to evaluate the efficacy of people's practices." Drawing on past work, Blomberg et al. (1993) describe four principles of ethnography: working in everyday settings, taking a holistic view, providing a descriptive account of events and activities, and taking a members' point of view.

Grounded Theory

Grounded theory, a method of inquiry for both collecting and analyzing data, supports an interpretive research approach for describing and classifying practices (Strauss and Corbin, 1998). The original intent of Strauss and Corbin (1998) with grounded theory was for inductively deriving theory. It is often used as a generative approach for identification of topics and issues (Charmaz, 2006). For transcriptions of interviews and other documents, a code set is developed that serves as an index into the text with codes describing the content of text segments. The original text and the index are subjected to further analysis including identification of concepts and categories related to participant language and actions. In addition, memos provide a formal mechanism for documenting field experiences and impressions as well as topics for narrative development. Memos, as a kind of field journaling, afford a time to recall, reflect, integrate, and conceptualize that draws on field experiences as well as field notes and transcription codes (Emerson et al., 2011; Creswell, 2007). Strauss and Corbin (1998) discuss the value of memoing in the field and describe the transition fostered by memos from collection of field data to working with concepts.

Situational Analysis

Clarke (2005) addresses larger areas of influence through focus on 'situations' that are analyzed with differing lenses. Situational analysis, described as a 'theory/methods package' by Clarke (2005), is an empirical approach to complex arenas of collaborative work. It provides a 'regrounding' of grounded theory within a context via cartographic visualization of key elements with maps that capture organizational, social, and positional arrangements (Clarke, 2009). Situational analysis is used to capture salient factors and perspectives that influence data and its

management at a site. As explained by Clarke (2005), situational analysis adds 'analytic thickness' to a study and complements the 'descriptive thickness' (Geertz, 1973) of ethnography.

Longitudinal ethnography

Longitudinal research involves continuing participation with a site that provides the time to acquire a broader view of activities and allows time for those involved to reflect on events. It is useful in investigating the development of processes and responses to change. Qualitative longitudinal studies are an approach that has the "ability to link across levels, micro to the macro, especially during periods of sustained and dramatic change" (Farrall, 2006). In this study, longitudinal research permits the following of activities carried out collectively by socially active participants and captures changes in participant understandings of data concepts and arrangements over time. The co-construction of timelines was used as an open-ended activity that captured and organized events. The timeline provided both a retrospective view and captured current events. The concise visual display of this information drew out participant experiences as well as memories and assessment of events. Timelines were augmented, streamlined, and subset for use in various situations.

Constructing the field

Ethnographic 'field sites' today may consist of multiple parts which necessitates attention to 'constructing the field' for an ethnographic study (Amit 2000; Marcus 1995; Falzon 2009). Hine (2007) discusses multi-sitedness: "The strength of the approach comes from a willingness to pursue connections rather than accepting field boundaries that might on first sight seem obvious." Monteiro et al. (2013) with an 'extended design' perspective aims beyond solely a localized view "to capture how workplace technologies can be shaped across multiple contexts and over extended periods of time". With improved communication technologies and increasing digital connectivity, Blomberg and Karasti (2013a) also describe ethnographic inquiries that reach beyond study of a single site. These sites may be physical places or virtual spaces. They may include a set of individual locations that are connected and may follow a single object-of-study across multiple locations whether it is people, things, metaphors, stories, or conflicts. Indeed, Beaulieu (2010) suggests generalizing ethnographic inquiry from its early focus on geographic colocation and immersion in community life to the more inclusive notion of co-

presence given distributed communities with their virtual sites of communication and activity. For collaboration in contemporary digital environments, Marcus (2007) speaks of 'collaborative imaginaries' in describing new kinds of design arrangements and stresses the changing sensitivities needed to adapt to the more complex objects of study resulting from mobility and multi-sitedness in ethnographic inquiries.

There are diverse locales where work with data occurs that may include task-specific work groups or data repositories. The data landscape may be considered disjoint stovepipes of data work or alternatively as a patchwork of loosely coupled efforts that sometimes self-organize by developing shared concepts, vocabularies, and registries. Galison proposed and adopted a patchwork perspective on what he called mesoscopic histories in discussions of the disunity of science**,** that is, of science as a collection of 'highly structured pieces' where each piece may have a different framework, conceptual scheme and/or paradigm (Galison, 1997). In a similar vein, a meta view of information systems found information systems tailored to differing situations and interacting or connected via loosely coupled practices (Berente, 2009).

In constructing the field, Beaulieu (2010) reconsiders the traditional role of primary informant, suggesting the use of 'infrastructural allies' in order to underscore participation of the researcher as a partner with the other participants rather than a more unidirectional flow of information from informant to ethnographer.

3.2 Study Design

The study design is organized into three sub-sections: data sources, designating primary case, and constructing the field. In designing this multi-case study, fieldwork was carried out at each case's project center in order to observe participants working with data and creating infrastructure as well as interacting with external repositories. The unit of analysis and object of study are defined below as follows:

> Unit of analysis – The scientific research project. Research project members engage with data specialists in collectively assembling and managing field-based project data and related materials. Project participants conduct field-based research and represent a community with social, technical, and information dimensions. Each case is comprised of one project.

Object of study – Project digital data work and repository arrangements associated with assembling and organizing project data and related materials.

Each case studied involved a trajectory of data and data work that started with field generation and moved to a local component and sometimes to a repository component.

3.2.1 Data sources

The cases are projects in the natural sciences where participants make, and record as data, field observations and measurements of earth and environmental phenomena. The kinds of data collected by project participants are diverse, ranging from small tables of numbers stored in spreadsheets to large data streams from automated instruments. The projects were selected opportunistically, although they were required to meet the following criteria: 1) active with ongoing field campaigns that generate heterogeneous data; 2) in a state of 'data management readiness', that is, having an active concern with addressing data management and a commitment to data sharing; 3) working with external data repositories or networks; 4) ready access to project members and documents for an extended period; and 5) funding has been available for the project for an extended time period.

For each project there was a data ally. References made below to 'data allies' (Beaulieu, 2010) serve as a reminder of mutual interests and learning as a joint pursuit in understanding the layers of practices and perspectives in complex, dynamic data work systems undergoing change. In each case there was a common element of what Marcus (2007) calls complicity between ethnographer and informant to understand the local situation of providing care for research data in the context of a larger set of forces.

An overview of the three projects in this study is given in Table 3.1 with a descriptive short name for each case together with my fieldwork timeframe and information about the project. The short name in the header is given in order to provide descriptive reminders of equal length for the cases as well as to provide short, memorable names. The actual names are given at the beginning of each case discussion in Chapter 4 since anonymity is not required. In each case there was ongoing support for established field platforms that were instrumented. For the environmental sites, the platform is a single geographically located place referred to ecologically

as a biome. The atmospheric science case, in contrast, centers on a mobile platform that supports joint field sampling at multiple sites. The layers of project context shown in Table 3.1 summarize the organization of the projects.

Table 3.1: Description of the three cases in this study

Project Characteristic	EcoPrairie	EcoRiver	AtmChem
Case Study Timeframe	2013-2015	2013-2016	2013-2014
Project Focus	Shortgrass steppe prairie ecology	Large river wetlands ecology	Atmospheric chemistry
Project Science	Environmental, interdisciplinary	Environmental, interdisciplinary	Atmospheric, interdisciplinary
Field Study Site	Single locale, experimental rangeland	Single locale, nature preserve	Single platform, diverse locations
Project Context	1. individual members 2. project office 3. network office	1. individual members 2. organizational offices 3. project office	1. individual members 2. science project office 3. organizational unit of center 4. archive unit of center

3.2.2 Designating a primary case

The EcoPrairie case involved a project studying ecological processes on a shortgrass prairie. EcoPrairie was selected as the primary case for this study because of its site access and long-term engagement with data management. Additional factors included the presence of a data ally and my familiarity with the long-standing network of which EcoPrairie was a member. Further, after more than three decades, the closing of the site made data management activities visible and prompted discussion by project participants about the data work as the termination of the data management and shut down of the data system were planned. Dialogue with the 'data ally' occurred over a three-year time period (2013-2015) that included the period of project closing (2013-2014). Field visits when in-person interviews were conducted during two intervals, first from January-May 2014 and then during the summer of 2015.

Opportunities to work with the two secondary cases, EcoRiver and AtmChem, also occurred in 2013. EcoRiver, a project that studied as well as managed restoration of river wetlands, began as the site was making its first inquiries into data management. EcoRiver's geographic proximity to my location facilitated site visits and the availability of several data allies contributed to its selection as a site. Having a second case set the stage for comparative

study of data management and repository arrangements at two independent locations, both in the environmental sciences. Experience with the pair of environmental cases provided a base from which to expand to include a third case, a center in the atmospheric sciences. This third case enabled cross-case comparison of data arrangements in circumstances that differed in terms of discipline, organization, sampling, and interactions with archives.

3.2.3 Constructing the field

A number of extensions were made in my definition of 'the field' for this investigation that enlarged the concept of a data system as a complex sociotechnical object in one location to that of a system comprised of elements across a number of arenas. Such a 'constructed field' enables study of contemporary data arrangements. For example, the definition of the unit of analysis was expanded in terms of data collection after participation and observation of data work revealed new information pertinent to project scope. First, my investigation of EcoPrairie initially targeted the time prior to and during closure of this long-term project. The collection of additional data during a subsequent visit represents a purposive sampling strategy decision. Analysis of the initial interviews suggested additional interviews of researchers who after the closing were continuing use of the data, working with new projects, and providing input to new data infrastructure plans for data sharing. As a result, I extended my fieldwork to include interviews a year after the site closed. These interviews proved useful in providing a broader range of project-affiliated participants, in gaining insights from participants who had time to reflect on the closing, and in interpreting interviews that took place prior to shut-down.

A second expansion in my unit of analysis occurred relating to the case boundaries of EcoRiver. I focused initially on the launch of a new field station, but observations and interviews revealed the extent to which work at the station intertwined with that of the group of researchers studying and sampling at the same preserve. My study broadened in response to the unanticipated collaborative nature of research at EcoRiver. Interest in addressing data management developed across preserve partners rather than being centered at the new field station that was my original entry point.

Third, my view of the atmospheric science case, AtmChem, expanded because of its location in the same organization as an external data archive, hereafter referred to as AtmDM. AtmDM served as one of several archives used by AtmChem so presented the opportunity to

study an external archive. While not developed as a separate, full case within the study, a profile and ethnographic details of AtmDM are included in Appendix E. AtmChem and AtmDM were two units within the same organization that at times represented a single data work configuration. One of the priorities of this study is to understand the relationship of projects with archives so AtmDM adds significantly to the AtmChem case. AtmChem and AtmDM worked together during field expeditions but upon return from the field, one carried out 'local' scientific work with the data while the other carried out the work of an archive.

3.3 Data Collection

Data collection is detailed in four sub-sections. First the Institutional Review Board (IRB) process is reported. This is followed by a discussion of the longitudinal approach and then of site access and invitations to participate at the site. The final sub-section describes the ethnographic fieldwork.

3.3.1 Institutional Review Board

An Institutional Review Board (IRB) approval was obtained from the University of Illinois at Urbana-Champaign before the start of data collection. I submitted a request in February 2013 and the study was approved in April 2013. The approval was for a one-year interval with annual renewals thereafter until it was closed and archived in February 2017. The consent forms asked about the degree of anonymity to be ensured for each individual with the following specific permissions: use of excerpts from recording transcripts, photography of work areas or individuals, identity of individuals and in later instruments organizational identity as well. The interview instruments submitted with the IRB are in Appendix A.

3.3.2 Longitudinal approach

My participation varied by case occurring over a period of four years for two cases and nine months for the third case. Interviews were carried out at strategic rather than regular intervals. I adapted my interviews and participant-observations to the rhythms of activity at each site. My sampling took into account the differing timeframes involved: a) the annual and seasonal cycles of field activities that often determined access to researchers' time; b) research

participants' engagement in time-bound "project cycles" as funded activities for defined intervals together with their more "regular" day-to-day work; and c) institutional circumstances that impact data management activities with longitudinal engagement. My aim was to capture both established and transitional forms of data work as well as to document project relations with other repositories.

3.3.3 Access and invitations

In all cases, I had open access for visits, observations, discussions, and interviews. There were institutional commitments to data work in all three cases so participants felt supported in working on joint projects and speaking freely. There were periods when I was at or living nearby each site. For the environmental sites the field sampled was nearby. For the atmospheric science case, fieldwork involved researchers at a science center that supported community use of aircraft to visit a variety of geographic locations both nationally and internationally. I developed relationships with participants through face-to-face meetings and time spent at project work areas.

In this study, each case began with an invitation from participants. Two invitations were issued in conjunction with major events – the first at the opening of a field station research award requiring data management participation and the second at the closing of a research project where the information manager expressed the value of input from an information science perspective during the shut down of their data system. The third invitation resulted from a joint data center and an information science school research study providing students of information an internship with interested individuals at the center (Mayernik et al., 2015). Joint activities and product co-development created a venue for appreciation of existing practices and situations as well as for identification and discussion of matters of concern (Carroll, 2014). Activities and products varied due to the differing levels of technological complexity and scope, lengths of my fieldwork, and my relations with data allies at each site. Stories of participation are documented at the end of each case in Appendices B to D. They provide examples of the activities and products.

3.3.4 Ethnographic fieldwork

Fieldwork for the three cases was carried out at intervals over a four-year period (2013-

2016). Work with each case proceeded separately. The majority of on-site fieldwork took place in the first two years although one week of additional in person interviews were carried out in July 2015 for the primary case to gain a fuller view of the information environment. My participation in working groups and with contacts began when I was at the project site and continued by telephone and video calls when I returned to my home institution. Further details on fieldwork and data collection are provided in the individual case profiles in Chapter 4 and the extended project materials in Appendices B to D.

Activities at each site included tours, laboratory visits, document collection, note taking, and meeting attendance. Ethnographic methods included participant observation, interviews, informal discussions, coproduction of products, and memo writing. Interview length varied from a half hour to two hours. Interviews were semi-structured. Questions were broad in order to allow participants to identify topics, activities, and events that they identified as important in their data work. The majority of interviews were transcribed, indexed, and coded. Additional materials for building a case included project-related notebooks, photographs, documents, and websites. I began development of the corpus of project-related materials while in the field through collection of project-related documentation, generation of field notes, and memoing.

Table 3.2 provides and overview of interviews and group meetings for the three cases. For EcoPrairie a total of 29 interviews included 11 interviews with the information manager who was the data ally. An additional 9 informal conversations were held with the data ally upon request in response to ongoing events. As one of the four members of a working group, I was a coparticipant with this group that met for 15 two-hour sessions to plan and implement data migration. A half-day tour included visits to the field station, visitor accommodations, and the shortgrass rangeland acreage.

For EcoRiver, a total of 14 interviews that included 4 interviews of two hours each with the field station director, a key data ally who was a researcher at the site. I attended a number of informal group session including a meeting held periodically by the partners. In addition, I made presentations at their annual science symposiums. A half-day tour of the EcoRiver acreage was arranged in order to see the field station and the configuration of land, lakes, and levees along the Illinois River. Finally there were 10 telephone meetings one to two hours in length as preplanning and debriefing for an EcoRiver data stewardship workshop.

For AtmChem and AtmDM (shown with an asterisk), there were a total of 31 interviews with 10 interviews of one to two hours each with an AtmDM data ally. I attended several AtmChem group meetings. Toward the end of my fieldwork, I held a wrap-up meeting with the AtmDM data ally. In addition, periodic discussions were held with a project scientist, a research data services specialist in the library unit at AtmCenter who served as my research mentor.

Table 3.2: Overview of interviews and meetings for the three cases

	EcoPrairie		EcoRiver		AtmChem/AtmDM*	
	# Individual Participants	# Interviews	# Individual Participants	# Interviews	# Individual Participants	# Interviews
Data Ally	1	11	1	4	1*	10
Data Manager	3	3	-	-	-	-
Researcher	6	6	4	4	11	12
Partner Researcher	2	2	-	-	-	-
Unit Manager	2	3	1	2	-	-
Project Staff	1	1	2	3	-	-
Resource Manager	1	1	1	1	-	-
Technical Staff	2	2	-	-	2 + 1*	3
Students	-	-	-	-	1	1
Sofware Engineer	-	-	-	-	4*	5
SubTotal	18	29	9	14	20	31
	# participants	# meetings	# participants	# meetings	# participants	# meetings
Working Group	3 to 4	15	3	10	-	-
Workshops	-	-	15-25	3	-	-
Informal Group Session	1 to 6	9	1 to 16	8	2 to 4	3
Science Partner Meetings	-	-	120	3	16	1
SubTotal	-	24	-	14	-	4

3.4 Analysis

Analysis took place over an extended period beginning at the field sites but continuing at a distance after interactions with the sites ended. I aimed with analysis to identify topics and develop sensitivity to issues as an incremental contribution to understanding data work in addition to contributing to 'theorizing' (Weick, 1995) or 'theory-making' (Czarniawska, 2005) that, in this study, relates to data-generating projects and repository configurations in the contemporary data landscape.

Memoing and situational mapping were key activities both when immersed at the site and while interacting with participants at a distance. Memos were frequently written after an interview or upon re-reading a transcript. A number of the interviews that were exceptionally information filled were transcribed fully or partially within a day or two of the interview so that

memos could be written. Thematic coding of these by hand with a focus on language use and data work informed ongoing work by highlighting issues and topics relating to data arrangements and responsibilities. Situational maps were developed on an ongoing basis. Mappings captured elements of interest in understanding the multiple contexts of a project involving intra-organizational groups, external organizational partners, domain forums, and communities of practice. With significantly different organizational arrangements in the three cases comprising this study, my aim with situational maps was to build awareness of the data work arenas.

After the first two years in the field, analysis became the primary focus. ATLAS.ti software was used for assembling and coding transcripts by case and as a single collection across all cases. Focused coding was carried out on selected core categories. In the iterative process of moving from fieldwork specifics to more general concepts, within case coding and analysis were carried out in integrative sessions aimed at considering groupings of code categories together with memos. Theme building was accomplished through reflection on and elaboration of existing codes as well as identification of new categories. Situational maps were created and reviewed in order to consider the data work within the full context of digital data arrangements.

In the fourth year of the study (2016), cross-case analysis was begun. This afforded a meta-view of commonalities in data work. Analysis provided deeper insights into the concepts of data pathways, data work, and pre-archive collections. A non-linear view of the initial analytic framework emerged together with an appreciation of the variety of configurations evident in practice that prioritized data work and science work differently.

3.5 Coparticipation

As a coparticipant at the sites, I contributed to activities that elicited and documented project related information. I worked together with participants to understand data situations at hand and sometimes brainstormed on potential actions. I brought to the discourse perspectives informed by my background in environmental sciences, information management, information sciences, and science studies. I gave priority to development of language and narratives. In particular I was attentive to discussions such as those concerned with community vocabulary, elements of design, forward planning, and retrospective reviews. I aimed to empower participants by prompting reflection that fostered conceptual and contextual awareness. My

contributions to discussions often aimed to lead participants to recall past events relating to current events in order to elicit 'tacit' or informal knowledge of an event or insight. In these venues, I was a proponent for joint documentation activities as a mechanism to prompt continuing interaction. The co-construction of products such as timelines, posters, talks, and technical reports by participants that included my participation as an ethnographic partner provided critical opportunities for clarification and co-learning throughout my engagement. Examples of collaborative products are listed as 'Stories of Participation' in Table 3.3 and included in the expanded case descriptions in Appendices B.2.6, C.2.4, and D.2.4. Participating collaboratively in activities with participants represented an opportunity to give back to the community, in accordance with the values of participatory design and ethnography.

Table 3.3: Stories of participation

Case	Stories of Participation
#1 EcoPrairie Appendix B.2.6	1. Developing a distributed or hybrid model 2. Data migration poster 3. Working group technical report 4. Project closing data activities list
#2 EcoRiver Appendix C.2.4	1. New field station meetings 2. Field station data management report 3. Annual meetings presentations 4. Data stewardship workshop
#3 AtmChem Appendix D.2.4	1. Briefing series talk on data curation 2. Data stewardship organization-wide poster

My participation is best described as multi-faceted. It was distinguished by an equal focus on a) capturing data work and data arrangements as an object of study and b) participating in and contributing to activities and co-learning where interaction was 'a resource for engagement and sociotechnical integration' (Gjefsen and Fisher, 2014). Although a coparticipant, my stake differed significantly from that of other participants. Whatever the outcome, there was something for me to report since success, failure, or any combination of them represents a finding. As participants, we were reflective practitioners (Schon, 1983) with a general interest in learning how the activity fit within broader contexts. The primary focus of other participants, however,

remained on the task at hand and its future, since they typically had a stake in particular outcomes. Keeping the story and its context in the forefront of my mind was an important aspect of my participation. One data ally mentioned my contribution as "a sounding board that can see the big picture" and identified participation in my ethnographic study in a project report as an 'impact on other disciplines'. In addition, as an outsider with translocal experience, I was able to distinguish some data work situations as local 'troubles' or as larger-scale 'issues' (Millerand et al., 2013).

3.6 Strengths, Limitations, and Robustness of the Study

There are limitations and strengths to ethnographic studies as well as concerns with robustness when conducting qualitative research. Limitations in this study included the quantity of data and data selection, time and unevenness of data collected, differences in settings and generalizability, and variations in my role. The strength of rich, descriptive fieldwork and analysis carries with it the generation of a preponderance of data that is not structured or intended for statistical study. A longitudinal study of multiple sites compounds the amount of data generated. Space constraints of the dissertation text required selection and compaction of material. On one hand my research data provided the scope needed to observe data arrangements; on the other hand, inclusion of only a limited number of stories tied to a coherent narrative is possible for a dissertation. In maintaining focus in this study on local data arenas in the natural sciences, not all topics that emerged during analysis were pursued. From the data available, selections were made based upon their effectiveness in conveying information about configurations relating to data work and to illustrate responses in practice relating to unresolved data issues. The focus of this study on local components rather than the sampling site or the archive represents both a strength and a limitation. Information on macro forces such as external major events and relations with partners are included in case timelines and nominally in the presentation of the case studies.

Balancing the research and participant roles was a continuing effort throughout. I addressed this challenge with frequent memoing after participation so the memo served as an observation, documentation, and analysis tool (Emerson et al., 2011). With three sites, the need to participate regularly represented a practical limitation. I conducted periodic member checks

with my data allies to discuss recent observations, activities, or emergent findings to clarify my understandings. In addition to member feedback, I prioritized making my information available to them, often through participation in joint activities, which could inform their everyday work. These periodic interactions were time-consuming. On-going analysis of the field data in concert with member checking informed subsequent fieldwork as the study progressed.

In addition to maintaining active engagement with participants, research engagement with the data takes time. Time as a limiting factor was addressed by designating one case as primary. With this strategy, a more limited set of materials was collected and analyzed for the two secondary cases. Though a continuing presence would be optimum, it was sometimes logistically difficult. In periods when fieldwork events overlapped, choices about participation were made with an eye to continuity of engagement. The scope of my study also led to difficult choices about what materials to include in the analysis. An adaptive strategy was adopted given the differing social, technical, organizational, political, and information situations of the cases.

In making sense of data work and repositories in everyday work settings using qualitative research methods, it is important to consider the appropriateness of the study process. Prolonged engagement and persistent observation coupled with multiple cases, triangulation, thick description, and member checking were central to this study. Such a variety of strategies contributes to the rigor and trustworthiness of qualitative research (Creswell, 2007). Discussions carried out during collaborative activities provided opportunities to clarify issues, validate understandings, and enrich interpretations in addition to creating occasions for participants to reflect upon processes, transformations, and actions. Further, use of a participatory design approach increases sensitivity to participant views and opportunities for engagement (Blomberg and Karasti, 2013a, 2013b; Susman and Evered, 1978). A three case design adds cross-site analytic strength to the study. The use of multiple research methods adds to the robustness of the study by providing multiple data sources that enable triangulation. Triangulation provides varied perspectives, situated credibility, and insight into critical issues (Creswell, 2007). Whether dealing with one or many cases, however, the diversity of technology, methods, practices, institutional arrangements, partnerships, and funding sources that impacts any one particular setting precludes simple generalization to other cases.

3.7 Summary

This chapter summarized the research design, provided background on research methods, and detailed the design of the study. The next chapter presents the three cases: EcoPrairie, EcoRiver, and AtmChem. In Chapter 4, each case is profiled with highlights of project data work. Participant voices are captured and presented throughout the following chapters and appendices. When quotes are included, attribution is made to the general role of the speaker. Participant categories included research scientist (RS), data manager (DM), project staff (PS), project management (PM), resource manager (RM), software engineer (SE), systems management technical staff (TS), and unit manager (UM).

CHAPTER 4. THE CASES

The three longitudinal ethnographic case studies about project data work configurations at scientific research sites located in the United States (see Table 3.1) are presented below. Figure 4.1 captures salient features of the projects. One case involves fieldwork with sampling at many locations, and two cases involve single-site sampling fieldwork. These field-based research projects investigate natural systems so contrast with laboratory-based research that is more controlled than is possible with the natural environment. Researchers in the field coordinate observations and measurements within a designated sampling area, generating project-related data that collectively has been characterized as 'interrelated heterogeneity' (Goodwin, 1995).

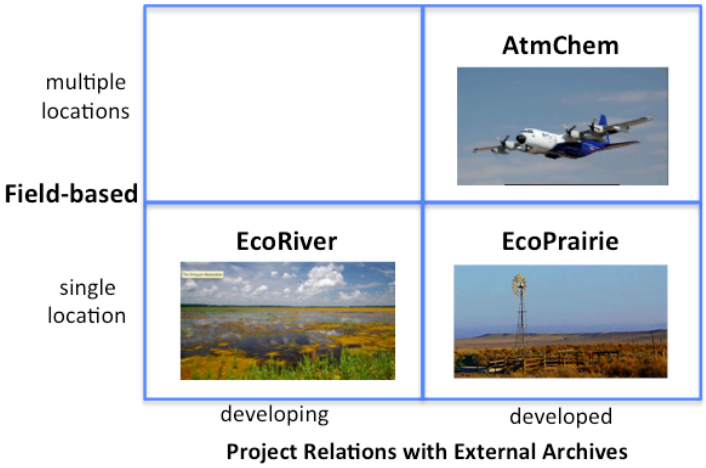

Figure 4.1: The three cases are shown in a four quadrant presentation that highlights their difference in relations with external archives (x-axis) and in sampling arrangements at one or multiple locations (y-axis).

Each case is described in two parts. The first part introduces the project, giving background on the project and its context. In the second part, three aspects of data work are described: the data management role, collective data work, and relations with partners. Collective data work across the cases involves local data management, systems support, data assembly, and a project website. An expanded report on data work and a timeline are provided in

an appendix for each case. Case timelines outline the larger context for current data activities. Expanded reports in Appendices B-D provide more in-depth coverage of the topics, often using the words of the participants in order that their voices are heard directly.

4.1 The EcoPrairie Case

4.1.1 EcoPrairie case overview

The overarching mission of this grassland project, located east of the Rocky Mountains and at the western edge of the Great Plains, is to understand the complex mix of organisms, properties, and processes that describe a prairie grassland and this ecosystem's response to disturbance (Lauenroth and Burke, 2008). The project, known as the Shortgrass Steppe Long Term Ecological Research (SGS LTER), will be referred to as EcoPrairie. EcoPrairie provides an example of project-based data management (Stafford et al., 2002) and its evolution. It illustrates long-term thinking about assembly, migration, and preservation of project data.

Table 4.1 summarizes the EcoPrairie case including a brief case description as well as my study timeframe and activities observed. A timeline of events (1939-2015; Appendix B.1) provides historical context. Three activities in particular are described: site closing, data work involved with shutting down the project data system, and the emergence of interest in designing for interoperability of systems using a distributed or 'hybrid' model.

Table 4.1: Overview of the EcoPrairie case

Case Name	Case Research	Case Description	Study Fieldwork	Study Activities Observed
EcoPrairie	Shortgrass Steppe prairie ecosystem	* Site designated in 1937 * EcoPrairie begins 1982 as LTER project * Project community co-located with fieldsite * Fieldwork at designated local area * Embedded data manager * Mature local infrastructure developed slowly	2013-2015	* Site closed in 2014 * Data migration * Distributed or hybrid model

EcoPrairie spanned thirty-two years (1982-2014) as a Long-Term Ecological Research (LTER) project funded by the National Science Foundation (NSF). Its project office was located on the Colorado State University (CSU) campus at Fort Collins. The LTER field sampling was carried out nearby at the Central Plains Experimental Rangeland (CPER) with a Prairie Grassland study site extending east of CPER. This land is owned by the United States

Department of Agriculture (USDA) Agricultural Research Service (ARS). Two field stations are situated on the CPER property: a USDA field station and an LTER field station established with support from the National Science Foundation and CSU. The CPER has a long history as a research site that can be described by four funding periods related to major research programs:

Time Periods

- Period 1: 1939-ongoing; USFS and USDA, ARS, Agricultural Experimental Station (ARS AES)
- Period 2: 1968-1974; NSF, International Biological Program, Grassland Biome Project (IBP GBP)
- Period 3: 1982-2014; NSF, Long-Term Ecological Research, Shortgrass Steppe Project (LTER EcoPrairie)
- Period 4: 2012-ongoing; USDA, ARS, Long-Term Agro-ecosystem Research Network (ARS LTAR)

Researchers from multiple sectors, including academic, government, and non-profit, worked together at the field site during these periods. Some contextual background follows that sets the stage for the story of this project's work with data.

The third period: EcoPrairie and its closing

This case focuses on collective work with data at the end of the EcoPrairie project (2012-2014) that brings Period 3 to a close. EcoPrairie funding for the LTER EcoPrairie period was contingent from the outset on taking responsibility for data management, particularly in light of scientific discussions of ecosystem time scales spanning decades to centuries and lessons learned about data work in the preceding IBP Grassland Biome period (Likens et al., 1987; Kingsland, 2005; Aronova et al., 2010; Gosz et al., 2010; Peters et al., 2014; Willig and Walker, 2016). Established as an LTER project in 1982, EcoPrairie carried out ecological fieldwork as a member of the LTER Network. The network began in 1980 with six location-based projects supported in a special NSF extended funding arrangement for 'the long term' (Callahan, 1984; Franklin et al., 1990; Hobbie, 2003). An LTER site is a network subprogram that is funded independently. The network grew by adding sites over the years reaching a total of twenty-six by 2004. Each LTER project was focused on the study of a geographic site's ecosystem known as a biome. In the LTER period, data collection was guided by network-wide themes and managed by each site in contrast to the earlier IBP approach in Period 2 when data procedures were

standardized and managed at centralized locations (Golley, 1993; Kingsland, 2005; Coleman, 2010).

The funding cycles of the LTER project in the third period with principal investigators, lead data manager, and academic organizational units at CSU are summarized in the Appendix Table B.2.1. Within these cycles, the project was led by a series of six principal investigators (PIs) together with their co-PIs while data management was led by a series of three data managers. Though the table shows the academic affiliation of principal investigators during this thirty-two year project, it does not convey the wide range of research partners affiliated with EcoPrairie and the CPER carrying out field and laboratory studies, data analysis and modeling as well as rangeland management. The EcoPrairie Project Office beginning with an affiliation at the CSU Natural Resource Ecology Laboratory (NREL) changed with the lead PI. After the third funding cycle, the lead investigator position together with the project and data management offices rotated to different university departments. The project returned after a twenty-year interval to NREL for the closing period.

NREL is a grant-funded, self-sustaining laboratory formed in 1969 as the project center for the IBP Grasslands Biome Project. Today, some of the research projects associated with this laboratory carry out research on design, development, and use of data systems. NREL maintains support for two laboratory information technologists and also makes use of university information technology (IT) services as well as library services that include digital collections and an institutional repository. A university-wide Information Science and Technology Center (ISTeC) was launched in 2002 that created a university-wide forum concerned with technology and digital content. The EcoPrairie information manager joined in ISTeC activities such as data management planning and surveying of the CSU digital environment. In 2008, the Vice President for IT at CSU was appointed interim Dean of Libraries and subsequently held a joint appointment as Vice President and Dean, thereby bridging information *and* technology.

When EcoPrairie's sixth LTER renewal proposal in 2008 was unsuccessful, the project was placed on probation and asked to submit a new proposal in 2010. This proposal was also unsuccessful so two years of funding (2012-2014) was provided to close the project. My fieldwork occurred during these closing years of the project when the data system was shut down and data was preserved in planned as well as unanticipated ways. The next section provides a summary of the data work and project-repository relations at EcoPrairie.

4.1.2 EcoPrairie project data work

The role of data management existed at EcoPrairie for the lifetime of the project. As an LTER network member, each site was required to designate a data manager. EcoPrairie started with a part-time data manager in 1982. The position was made full time in 2008 and called by that time 'information manager'. Data support developed continuously for the project until 2011 when project funding difficulties began and data management support decreased until the site closed in 2014. From 2008 to 2011 the information manager oversaw a small team of part-time members with the aim of coordinating the assembly of data from project members. Data work included workflow design, data entry, data assembly, geographic information systems (GIS), support for project partners as well as design and population of a project website. Data work also involved participating in community efforts within the LTER network-wide Network Information System (NIS). EcoPrairie membership in an LTER-wide Information Management Committee (IMC) afforded ongoing learning and leadership opportunities. The IMC was a community of practice where members were engaged in development of metadata standards, community best practices, and community decision-making (Appendix F).

Project data work involved establishing and promoting the assembly and coordination of data. A data workflow for collective data assembly was developed that became known locally as the project data system that in essence was a data repository. Data management was proactive in developing ties to the field by providing services such as online sampling permit applications and field notebooks filled with field sampling information for reference. Conventions for documenting variable names and units were developed and then coordinated with an LTER Network unit registry. With technical support provided by campus-wide services, an EcoPrairie website was established. It was considered the norm for LTER member sites to maintain websites providing a variety of project-related materials that not only made information about the project available to the public but also established an information environment that fostered communication among project participants as they negotiated the items that best represented the project as well as what had been learned about the biome.

Even before the LTER program focused on long term data from site-based natural systems, EcoPrairie's location at a land-grant college with its mission to foster the applied use of knowledge set the stage for working in partnership with academic, government, and business sectors as well as professional associations. LTER working in collaboration with scientists at the

USDA/ARS and other CPER partners, created a critical mass of researchers with overlapping interests at a shared location.

The closing period

Closing activities for the site focused largely on ensuring data were ingested into the LTER network system rather than on loss of local data and information management. The aging hardware and software were left in place with little attention since they were not compatible with other on-campus research data systems. In a final two-year supplement proposal, the information manager requested support not only for delivery of data to the NIS, but also for support to document the legacy of the site, to update metadata to meet the evolving community metadata standard, and to continue to participate in the larger information management context. During the closing years, the information manager's work broadened within the department to include data management planning for new projects, work with other project data systems, training or supporting graduate student work with data, and participation in emerging cross-campus data efforts. At project end, the information manager took a new position outside the university (see Appendices B.2.4, G.2, and G.3).

During the closing period, continuing changes in data work were evident. Data work did not wind down but instead involved actively addressing the update of metadata for project core datasets in order to meet ongoing updates to the Ecological Metadata Language (EML). Work with the LTER network system also involved gaining experience with a recently deployed metadata validation application developed in conjunction with the LTER IMC. The work involved not only fixing unanticipated incompatibilities in database functionality but also update of local metadata conventions to include unique identifiers tied to documentation of partners and monitoring of rapidly developing data citation options.

Due to inquiries made by the project information manager and a chance encounter with a an archive finding guide, work with the library on digitization of old reports and photographs led to formation of a working group to prototype migration of EcoPrairie data to the library-supported institutional repository. Though the digital asset management system supporting the repository would be replaced after EcoPrairie closed, the project data and other project-related artifacts were migrated to the library system as a pilot project. This effort extended the institutional repository functionality to include scientific research data. It required expanding

existing library metadata practices to include metadata fields critical to data and project description. The organizing and structuring of data augmented the system originally designed for browsing to include data download and harvest. This evolved into what was recognized as formation of a project collection that included not only the data from the project data repository but also related project artifacts such as species lists, field notebooks, photographs, and project proposals.

The availability of the data in a new, nearby university-based data repository led to conceptualizing a hybrid model that involved the notion of working with systems to access and perhaps visualize the data now accessible in the academic library system. This was discussed within NREL, the department administering the EcoPrairie project office. This kind of configuring of systems required establishment of shared specifications to support access to datasets within a collection for automatic harvesting by other systems. This design feature was not available at the time for project data in the library system.

The extended closing period of the EcoPrairie project provided the time needed to package and migrate data as well as to explore options available at closing and to identify as yet unrecognized data issues that contribute to the development of data-related concepts and products. Thus project data moved to two different repositories, the LTER network system storing datasets queriable at the variable level and the library repository storing a project collection that included the data. The question arose as to whether the relationship between these two repositories should be described in the metadata associated with the unique identifiers assigned to the digital objects. Inquiries were made but relating repositories and considering concepts such as primary or secondary sources were not topics that those involved were ready to consider.

A more detailed account of the case that includes the voices of participants is provided in Appendix B. Four stories about my participation in activities are discussed in detail at the end of the expanded case study report (Appendix B.2.6). They involved the naming of an emergent concept of distributed work within a network of repositories, support in developing of a data migration poster to present at a conference of data professionals that was attended by three of the four working group members, writing a technical report to document the effort in transforming and migrating data from a scientific research project to an institutional repository, and developing a project closing data activities list.

4.2 The EcoRiver Case

4.2.1 EcoRiver case overview

At this wetlands preserve on the Illinois River, researchers carry out ecological studies and restoration science in nearby lakes, drainage areas, wetlands, and terrestrial sites as well as on the river as part of a floodplain river ecosystem (Sparks, 1995; Walk et al., in prep). The ecosystem is of significance because of its large watershed and its productivity (Sparks, 1992, 2010). The preserve consists of approximately 7,500 acres near the town of Havana, Illinois. A system of levees built by the Army Corps of Engineers is an integral part of management of the lakes and reservoirs on the property along the river. This project, located on land purchased by the Nature Conservancy and called Emiquon Preserve, will be referred to as EcoRiver. The Nature Conservancy, that has an onsite office and a nearby regional office, will be referred to as EcoRiverOrg for this study. Three field stations are in the area. Two of the nearby field stations monitor and study wildlife in the area. These two field stations are part of the Illinois Natural History Survey (INHS), a division of the Prairie Research Institute at the University of Illinois at Urbana-Champaign. The third field station is Therkildsen Field Station at Emiquon, hereafter referred to as EcoRiver Field Station. EcoRiver illustrates data management planning spurred by funding agency requirements for data sharing. The major units in this case include

- EcoRiver – the partners involved with the Emiquon preserve
- EcoRiverOrg – The Nature Conservancy Emiquon preserve owner/manager
- EcoRiver Field Station – Therkildsen Field Station at Emiquon

The EcoRiver case provides an example of the introduction of the concept of collective data management within an organization and the community made up of multi-sector partnerships sustained over time.

Table 4.2 gives a summary of EcoRiver including a brief description of the case as well as the timeframe of my study fieldwork and some of the major activities observed during the fieldwork.

Table 4.2: Overview of the EcoRiver case

Case Name	Case Research	Case Description	Study Fieldwork	Study Activities Observed
EcoRiver	Large river nature preserve	* Site designated in 1858 * EcoRiver begins 2000 as preserve * Project partners co-located with fieldsite * Fieldwork at designated local area * No designated data manager * Nascent local data infrastructure	2013-2016	* Field station workshops 2012 * Collective flood study 2013 * Data management planning

A timeline of site events (1858-2016; Appendix C.1) provides historical context. The EcoRiver history is divided for discussion into the following periods:

Time Periods
Period 1: 1858-2000, DOI/INHS Field stations established
Period 2: 1980-1986, NSF/LTER Illinois Large River project
Period 3: 1986-ongoing, DOI/USGS/LTRMP and UMESC
Period 4: 2000-ongoing, The Nature Conservancy EcoRiver conservation plan
Period 5: 2008-ongoing, Field stations, floods, and management efforts

Over these time periods, a community of participants affiliated with differing organizations including university-based field stations, state and federal government agencies, as well as non-profit and for-profit sectors have monitored and conducted research at the preserve. Activities at the preserve range across biomes from river to wetlands and terrestrial, across disciplines from anthropology to biology, ecology, and aquatic environments, and across a variety of sectors involving research and education outreach. One of the many partners explained some of the challenge introduced by the complexity of this large-scale endeavor: "You have to think at multiple scales …". Applied research includes field monitoring, process studies, intervention prototyping, modeling, and floodplain restoration. Participants from EcoRiver collaborate with colleagues at the Department of Fish and Wildlife, the U.S. Army Corps of Engineers, the U.S. Geological Survey and Dickson Mounds Museum among others. A report about the initial data management effort at the EcoRiver Field Station contains further information on EcoRiver partners (Baker, 2016, Table 2).

The fourth and fifth period: A conservation plan and data management planning

EcoRiver drew on state-supported natural history survey work that began in the 19th

century (Mills, 1958; Hays, 1980; Bocking, 1990) anchored at field stations (Period 1) and with an interdisciplinary LTER project focusing on long-term study of ecological biomes in Period 2. The Illinois Large River Long Term Ecological Research Project (1980-1986) contributed to conceptual development and vision for studies in subsequent periods particularly with respect to the theme of disturbance. The Illinois River LTER together with the Illinois Water Survey sponsored an early LTER data management workshop (Sinclair, 1983). With research focus on time-series data complimenting monitoring efforts (Swanson and Sparks, 1990), there followed some foundational papers after the site was terminated on the flood pulse concept that recognized disturbance as creating patterns of ecosystem development (Junk et al., 1989; Sparks et al., 1990; Sparks 1992). This work continued under the Environmental Management Program in 1986 for the Upper Mississippi River System (Sparks, 1992, 2010) begun in Period 3 with federal funds to support restoration projects (Walk et al., 2016). During Period 4, EcoRiverOrg created the EcoRiver restoration project. In Period 5, the INHS field stations were moved to UIUC management. These events established not only an academic location for ecological research but also a collaborative endeavor particularly suited for the study of a coupled human-natural system. The set of partners assembled to support scientific management at EcoRiver bridged many divides including those between basic and applied research as well as those between land management and environmental policy.

The EcoRiver project conservation plan in 2000 was followed in Period 5 by construction of a new field station and data management plans incorporated into two successful research proposals. To prompt further research at the preserve, the EcoRiverOrg statewide science office issued an invitation to the University of Illinois at Springfield (UIS) to build a new field station at the preserve. With field-based experience recognized as critical to support for ecological research and teaching (e.g. Billick, 2010), UIS seized the opportunity in 2008 to establish the EcoRiver Field Station. Data management was incorporated into the field station development process as a result of an NSF field station planning grant awarded in March 2012. I was one of two information managers invited to represent data management at the field station's planning workshop. A 'flood of the century' for the Illinois River occurred soon after in April 2013. A second data management plan was written for a targeted research study proposal that was funded after the river overtopped levees and reconnected lakes and reservoirs to the river. The two data management plans prompted awareness and stirred community interest in management of data.

The plans represent the initial state of participant views at the outset of EcoRiver data management. Data work and project-repository relations at EcoRiver are described in the following section.

4.2.2 EcoRiver project data work

Interest in collective data management at EcoRiver developed simultaneously at the field stations and EcoRiverOrg. Data at the time was held in disparate personal, station, and institutional assemblies without overarching coordination or easy access. Researchers including technicians and other staff working on or near the preserve carried out data work. The idea of a separate position for data management was absent. Yet, increasing amounts of data together with past experience with data loss combined to create receptiveness to the concept of data management as a significant issue. In this collaborative research environment, there was a readiness to consider options and to move from the notion of spreadsheets to new approaches for collective data work.

Participants initially were optimistic about collective data work, imagining analogies with library organization of print materials. In planning discussions, there were embedded assumptions about the availability of standards, protocols, and techniques and the ease of their enactment. A lack of exposure to the new language of data management and data curation associated with data-related processes led to explanations of their data needs in oversimplified terms involving solutions assumed provided by technologies and databases. Ambiguity in what constituted a dataset and in understanding how data would reach the web underscored a lack of local data conventions and practices. In time, however, EcoRiver expanded upon the concept of data management to embrace that of data stewardship because of their familiarity in ecology and conservation with the concept of land stewardship.

When funding permitted, several partners joined together in what was subsequently considered a pilot project in collective data management. This first experience with assembling data in an open manner supported by computers uncovered data issues including the need for external technical support such as that available at the campus level. Those involved in assembling the data developed a staging arena with three versions of the data that ultimately were ingested into a database on a standalone computer. They and the researchers working with them found the naming and reformatting process for data more time consuming than imagined.

They were surprised by difficulties with database synchronization. These 'hands on' experiences were invaluable in broadening understanding of the number of interconnected issues involved in data work and in increasing time estimates when planning collective data work. The pilot effort to assemble data involved part-time staff and students, a university laptop with a non-relational database application, cloud storage, and data upload from various participants that ended with a meeting with university-wide IT support to explore online access and delivery.

Basic computer systems support for EcoRiverOrg existed formally as a statewide, organizational administrative management effort while basic systems administration also existed at the field stations. The priority given to Internet connectivity was evident from the rapid connecting of the new field station computers to the Internet but plans to develop workflows for data within UIS or EcoPrairie were absent. Further, EcoPrairie technical support was distributed across organizational entities such as the campus-wide IT that hosted servers for some units while partner field stations had ties to an institute at UIUC. The library at UIUC where development of a data repository was launched in 2016 supported an institutional repository for UIS. A website integrating the work of EcoRiver partners had been established and continued to be supported by a museum partner. Its broad scope included information on site history, events, recreation, volunteer efforts, and visitor services as well as a research page. Early enthusiasm at EcoRiver with the potential for social media and data display, faded as these became recognized as time intensive undertakings.

Many individual researchers were grappling with assembling and organizing their own data on laboratory computers sometimes combining different streams of data and sometimes developing different processes for various groups of data. The transition to a template approach emerged but did not mature. During this time prior to in-depth experience with collective data management, some participants were able to articulate the concept of a data repository as an 'essential place' that encompassed many stages of development and handled different kinds of data as well as serving as a 'depository'.

Relationships with partners varied in terms of their scope and topics of interest. Partnerships involved a complex mix of science and data efforts that had evolved over time to support decision-making about use of natural habitats. Large-scale data efforts with government agencies involved multi-state partnerships and federal supervision of funds. In 2016 a data workshop was held that was mindfully inclusive of all partners. Data vocabulary including

dataset, data system, and data repository was introduced together with the notion of a 'Minimum Data Team'. Participants intermixed considerations of individual, EcoPrairie, other regions, and statewide efforts since existing coordination existed with all of them.

A more detailed account of the case including the voices of participants is provided in Appendix C. Four stories of participation are presented at the end of the expanded case study report in Appendix C.2.4. The initial field station planning grant meeting in 2012 gave data managers a voice at the table for the first time at the site. Two data representatives coordinated their suggestions: 1) recognize data management by designating a data manager and 2) begin collective data work with the EcoRiver sampling permit process not to control but to document. The second story provides details on a data management report about a study of the new field station at EcoPrairie providing information on its long history and the themes of data management, data sharing, and infrastructure soon after the station's opening. The third story describes the co-construction of posters by small groups and myself that prompted dialogue as well as and developed a data management presence at the annual science meetings. Finally, a fourth story presents details on the data stewardship held in 2016 that presented an information management job description that was shared with participants and is summarized in a report that was deposited in the institutional repository (Walk et al., 2016).

4.3 The AtmChem Case

4.3.1 AtmChem case overview

The Atmospheric Chemistry Division (ACD, subsequently renamed Atmospheric Chemistry Observations and Modeling Division, ACOM) is a division within the National Center for Atmospheric Research (NCAR, 2012; NRC, 2007). NCAR is managed by the University Center for Atmospheric Research (UCAR), an independent non-profit organization with a consortium of more than 100 members. For instance, nearby University of Colorado at Boulder in Colorado is one of the members. At the time of this study, ACD was one of three divisions of the NCAR Earth System Laboratory (NESL). It focused on advanced atmospheric chemistry measurement and modeling capabilities. Also at NCAR, the Data Management Group (DMG) provided data services for atmospheric science projects across the organization as part of

the Computing, Data and Software Facility (CDS) within the Earth Observing Laboratory (EOL). This case involved three units that are referred to as follows:

- AtmCenter – National Center for Atmospheric Research (NCAR)
- AtmChem – NCAR Atmospheric Chemistry Division (NCAR ACD)
- AtmDM – NCAR Data Management Group (NCAR DMG)

In short, AtmChem and AtmDM are situated within AtmCenter, and AtmChem often makes use of AtmDM data services. The case provides an example of intra-organizational data arrangements in addition to work with multiple archives and modeling efforts with national and international atmospheric science community partners.

AtmCenter with core support from the National Science Foundation, is one of more than 40 Federally Funded Research and Development Centers (FFRDC; Hruby et al., 2011). National science and data centers are formal organizations that institutionalize the management and creation of research data (Mayernik et al., 2014). AtmCenter was launched in order to enhance capabilities across the atmospheric science community. It began research and data collection in the 1960s. The conduct of science by a national center differs significantly from that of individual research projects and laboratory groups in academic institutions. The organizational mission includes support for research efforts with size, duration, equipment, and technical needs beyond the scope of a single academic institution. AtmCenter is comprised of groups within laboratories and divisions. Some of the organizational units provide data tools and services within a laboratory or a division and some provide support across many organizational units. Technical support may involve field instruments, computer systems, data systems, gateways, and grids with their attendant hardware, software, and project data liaison activities. Increasing demands on these national centers in the 21st century by the communities served prompted national studies in 2003 and in 2016 (NRC, 2003; NAP, 2016).

Table 4.3 summarizes AtmChem including a brief description of the case as well as my study fieldwork timeframe and activities observed. AtmChem researchers participate in projects with national and international field campaigns typically supported by aircraft instrumented for multiple measurements. In addition to periodic field campaigns and data generation, AtmChem activities include continuing data analysis, integration, synthesis, and scholarly publication. It is a hub of collaboration for atmospheric researchers as research on chemistry impacts on climate

and air quality is conducted through multi-institutional field campaigns, laboratory experiments, and computational modeling.

Table 4.3: Overview of the AtmChem case

Case Name	Case Research	Case Description	Study Fieldwork	Study Activities Observed
AtmChem	Atmospheric chemistry	* AtmCenter planning begins in 1957 * AtmChem-like organizational unit begins 1966 * Division conducts multiple field projects * Fieldwork global at multiple remote sites * No designated data manager within unit * Expanding local data infrastructure; mature external archive relations	2013-2014	* Lab reorganization ongoing * Data catalog development * Multi-archive relations

A timeline for the project (1957-2016) provides an overview of the case (Appendix D.1). The AtmChem history is divided for purposes of discussion into the following periods:

Time Periods
Period 1: 1957-1966, Early AtmCenter planning and operation
Period 2: 1966-2003, Atmospheric chemistry identified as a key program
Period 3: 2004-2010, AtmDM established with an expanded mission
Period 4: 2011-2016, Further organizing for collaborative science

The fourth period

My fieldwork occurred during Period 4 in the midst of a multi-year planning period underway in response to an NSF inquiry about AtmChem and a national review of atmospheric chemistry research (NAP, 2016). These external efforts added to discussions during the generation of NCAR's traditional generation of annual reports (e.g. NCAR, 2014a) as well as a five-year strategic plan (e.g. NCAR 2014b) that provided an integrative look across the organization. Individual AtmChem project participants moved data to a variety of data archives external to their division. Project-level support for AtmChem fieldwork and data archiving was provided frequently either within the organization by AtmDM or by a National Aeronautics and Space Administration (NASA) data archive. Fieldwork duration varied from days to longer periods with repeat field visits. From 2005-2014, there were a total of 39 AtmChem campaigns that used an archive for an average of 3.5 campaigns per year. Of these, approximately 38% were archived at AtmDM and 36% at NASA. Data work and project-repository relations at AtmChem are described in the following section.

4.3.2 AtmChem project data work

AtmChem researchers largely managed their own data with support from a local systems administration group in their organizational unit. With much of the AtmChem data processed and analyzed by individual instrument-specific teams that generated data in-situ and submitted it to archives, data management did not exist as a separate category of work for AtmChem research groups. Further, given the diversity of instruments and data involved, a number of domain data standards were used. Advanced data handling capabilities for individual AtmChem researchers were supported by division-level and organization-wide technical services. A local systems administration group supported state of the art file sharing among project participants located at AtmChem in addition to services that were rapid responses tailored to requests from individuals or groups of individuals. Further, system administrators for research units participated in an informal intra-organizational community of practice that coordinated work across the organization and sometimes viewed projects in one group as pilot studies of interest to the rest of the community.

Project systems support by the AtmDM archive staff began with aircraft as shared platforms for reaching and sampling the field. The details of collective data work were negotiated prior to each set of field campaigns. AtmDM technical staff, often with the title 'software engineer', provided a set of services designed to accommodate the many field projects they supported. Data arrangements were made by AtmChem science projects with this or other established data archives during the planning stage of a set of field campaigns. Pre-deployment activities might include creating composites of the sampling site prior to a campaign, arranging for data assembly in the field, as well as initiating the process for archiving versions of the data and related project artifacts. Later stages of support involved assembly and access to three versions of the same data that were designated field, preliminary, and final. The archive limited data sharing among project participants for the first two versions while the final data were made publically available. Assistance was made available to bridge the research unit and archive unit at AtmCenter by an AtmDM staff member called the 'designated project data management contact'. Assistance could include help in packaging data for submission or in gaining access to data.

The AtmChem website was supported by a web specialist in the unit's systems administration group and at AtmDM by a web designer working with a metadata-driven system

of data delivery to the web. The AtmChem website was undergoing significant changes in representing its work by building a complete local to convey breadth of the unit's work. This was needed because of their submissions of data to differing archives such as a NASA archive as well as AtmDM, an archive within AtmCenter. The AtmChem catalog illustrated the difficulty of gathering the digital 'pieces' of the project together. Links pointed to local data storage, to science-driven websites, and to archive-driven websites, each playing an important role in the data work of science.

During my study, support for the atmospheric chemistry community in general as well as the role of AtmChem was under review (NAP, 2016). Ultimately, the review resulted in an improved articulation of the unique integrative blend of fieldwork and laboratory observations, modeling, and services being provided to university researchers.

A more detailed account of the case including the voices of participants is provided in Appendix D. Two stories of participation are described at the end of the expanded case study report in Appendix D.2.4. The first is a data management presentation I made to AtmChem as part of their biweekly AtmChem Briefing Series Forum where two major points were underscored. First, given their distributed archiving practices, I presented a survey of their data as it was presented on their website and discussed the value of an online local data catalog coordinated with and using the language of their partner archives. Second, I introduced the concept of data preservation as a new kind of long-term data work that differed significantly from the tradition data work familiar to research scientists (Baker and Millerand, 2010; Baker et al, 2013).

CHAPTER 5. DATA WORK CONFIGURATIONS

This chapter describes the analytic framework applied in my analysis of data work configurations. Following the basic framing presented in 5.1, section 5.2 discusses elements of data infrastructure supporting data work in the Local Component drawing on specifics from the cases. Models are developed for two of the eight elements identified: 1) the internal relations of science and data work and 2) the data work configurations of the three case studies. In the final section on findings, the elements of collective data infrastructure for data work of each case are summarized in Table 5.4. In addition, some trade-offs associated with managing data infrastructure are identified and explored.

5.1 Analytic Framework for Project Data Work

The analytic framework used to investigate project data work is comprised of three components: Field, Local, and Archive. As a conceptual tool, the framework ties together the work associated with the movement of data from its origin where it is first generated at sampling locations (the field), then analyzed and assembled for a purpose identified by the project (the local), and sometimes deposited in a data facility (the archive). This 'following the data' is an extended design inclusive of work with artifacts in the Field to the Local component including relations with the Archive Component.

Table 5.1 provides a summary of the characteristics for each of the field-local-archive framework components for the primary case. For the Field Component, a sampling plan takes into account the project scope and addresses field logistics such as the kinds of sampling, sampling locations, and expected field conditions. The Local Component involves individual researchers working independently as well as collectively. Collective data work carried out within the project Local component is referred to as the Local Data Collective. The Local Data Collective is an ensemble of social, technical, and informational arrangements that together constitute a mesoscale 'way station' between the field and their archive partners. The Archive component involves an archive external to the Local Component providing services for preservation and reuse of data to one or more projects. Data assembly starts in the field with notebooks, instruments, and project artifacts. In the field and locally, digital data resides on

various computers in many formats and groupings. A shared data system anchors a project information environment while an archive component includes highly structured data systems together with data access for many projects via a public website.

Table 5.1: Characteristics of collective data work for the framework components

Component Characteristic	Field	Local	Archive
Focus	sampling	knowledge making	data preservation
Scope	sampling plan	project plan	services policy for multiple projects
Data work focus	collecting data	using data	data archiving & access
Data assembly tools	notebooks, instruments, project artifacts, shared platforms	computer systems, data systems, website	archive systems, website
Digital data objects	data files and other study documents	data files, datasets, data packages, pre-archive collection	dataset collections, project collections

5.2 Elements of Data Infrastructure in the Local Component

The salient elements of data infrastructure for the local component that emerged from analysis of the cases in this study are discussed in the sections that follow. At a time when the Internet and related tools were developing rapidly (1991-1994), Star and Ruhleder (1996) identified eight dimensions of infrastructure associated with the development of a collaborative system for a distributed community of users. These are presented in Table 5.2 to consider with those developed below for data infrastructure associated with a local component.

Table 5.2: Dimensions of Infrastructure

#	Dimensions of Infrastructure Star & Ruhleder 1996
1	Learned as part of membership
2	Built on installed base
3	Embodiment of standards
4	Transparency
5	Embeddedness
6	Reach or scope
7	Links with conventions of practice
8	Becomes visible upon breakdown

5.2.1 Data management

The three cases in this study differ significantly in what today is called data management arrangements. EcoPrairie pursued collective data management and data work practices supporting a loosely structured, local project data system. EcoRiver was in the process of developing a concept of collective data management while AtmChem had available experienced systems management support group as well as advanced modeling capabilities.

EcoPrairie provides an example of how continuity of a data management position enabled the growth of local data infrastructure. In alignment with LTER network policy, an EcoPrairie information management position was funded by the project to facilitate the movement of data from the personal laboratory spaces of individual researchers to a local project assembly center that provided access to the data via the web. The data manager served as a coordinator for data documentation, data assembly, data systems development, and the website.

EcoRiver's efforts with collective data work were nascent. The approach taken was to launch data management planning as a site-wide activity that included developing a job description for a data management position (Appendix G.3). The role of a data manager was imagined as addressing many of the changes project researchers were finding needed attention – data assembly, data workflows, data systems, and data products. Researchers realized their own data practices would need to be modified or augmented to support collective data work. Descriptions of data management were expressed as "establishing a good framework", "phasing it in", and "all using the same template to make it easy for people to submit their data.'' With various levels of interest in data management, their approach anticipated change.

There was support for the collaborative data work of AtmChem project researchers provided by their division and by external archives including AtmDM within their own organization. One AtmChem researcher trained in atmospheric chemistry and modeling was well recognized for merging data from many project sources to create integrated data products. This researcher, referred to as the individual in the department with the most expertise in data management, stated emphatically several times in interviews "I am not a data manager". The statement indicates an awareness of data management as a role that differs from that of an atmospheric chemist. At the AtmDM archive, a software engineer commented on some difficulties that had been encountered in getting traction with project scientists on data

management issues:

> PIs, they are thinking very tunnel vision, they are thinking about the ongoing project. And they are thinking about their own instrument, or their own part of the project. And very few of them see the bigger picture. And so when I talk about data management, sometimes I get the 'deer in the headlights' look. (UM)

At AtmCenter individual field scientists generally managed their own data so collective data management was not defined as a distinct kind of work. Technical staff often had the title 'research assistant' when working in science-driven laboratories and 'software engineer' when working in laboratories supporting large data systems. The title 'research assistant' was used in referring to work both with technical systems a well as with data analysis and scientific knowledge making.

5.2.2 Systems management

Support for systems management is the second element that characterizes data infrastructure for the local component. Systems management was often referred to as systems administration or sysadmin in practice. EcoPrairie provided an example of a data manager who served as a coordinator of systems management for the project. Technical systems coordinated included the project data system, GIS work, the website supported by university IT web hosting services, and consultation by two technical specialists in the department of the project office. In contrast to these project-coordinated efforts, EcoRiver did not have a project-wide system management support group. EcoRiver project participants worked independently on computers with a variety of specialized applications.

A small AtmChem systems management group was responsible for developing and maintaining technical support for the individual AtmChem project researcher's work with software and data files. The group managed shared systems that ensured data files could be moved easily among investigators within the laboratory as well as externally to the atmospheric community outside the laboratory. With science driving data management, an AtmChem technologist observed that within organizational units such as AtmChem, it is the individual scientist who is the 'data architect':

> People tend to kind of solve, work on their data problems independently. They come to us when they are looking for technical infrastructure to support their solutions. We don't

> really do a lot of architecting solutions for data except kind of general infrastructures separate from the data … so right now the people designing for data are the individual scientists that are dealing with their instrument or their field project or their model. (TS)

5.2.3 Data assembly and data systems

A third basic element of data infrastructure involves data assembly and data systems. Across the three cases, language to describe data assembly for projects was in development, evolving as new kinds of data arrangements were made. Issues that arose in referring to data assembly are considered in order to convey the emergent state of understanding of collective data work and data infrastructure development. Legacy effects and situated understandings influence the vocabulary in use – data system, dataset, database – and often hamper communication. This study explored details of data assembly practices for EcoPrairie with a data system, for EcoRiver with researchers managing individual datasets, and for AtmChem making use of a variety of data archive services including support in the field.

EcoPrairie project researchers, in a manner observed throughout the LTER network, assembled data collectively using manual and digital processes. As a community proactive with data management and data sharing, use of 'data system' began in the decades preceding data access mandates (Holdren, 2013). As parts of the data workflow became digital, the assembled data was referred to as 'the data system' or 'an information system'. A heterogeneity of actions, interactions, and tasks carried out by the information management team, crafted to accommodate individual investigators, constituted and supported the system. The data system was not referred to as a data repository, its position as a process within the project and its academic institution seems to have precluded its recognition as an organizational entity.

In the absence of local conventions across EcoRiver, the grouping and formatting of data were particular to individual researchers and field stations. In references to data assembly and exchange at EcoRiver, researchers were at a stage of working on understanding 'what' was to be created for sharing. The use of the term 'dataset' was common in referring to a set of variables collected into a single data file. Single dataset files might contain data collected during one field visit or an assembly of observations gathered over time. Further, a mix of kinds of variables – physical, chemical, and biological – were assembled in sets of spreadsheets that included calculations. In considering data to share, investigators discussed ways of standardizing the packaging of their various datasets prior to deciding to initiate a database effort.

Data assembly at AtmChem was evident in collective work with merge files and modeling rather than with a local data system. Data storage systems, supported by system administrators at AtmChem, served as a place for AtmChem researchers to assemble and share instrument and model data. Data storage had traditional backups performed though bit deterioration, data or software migration, metadata validation, and data discovery were not within the purview of this work. A significant difference in understanding of long-term storage was evident. Systems administrators considered local file system storage a temporary holding facility for data files while some researchers considered the data as 'saved'. As a result, archive personnel were concerned about these data as potential 'orphan data'. This is an example of the kinds of misunderstandings found in today's period of transition when science projects are faced with enacting new practices involving long-term data preservation and public access.

5.2.4 Project websites

A fourth basic element characterizing data infrastructure for collective data work is a project website that contributes to a project's information environment. A website is a mechanism for identity building as well as community building. The gathered information, catalog creation, and content posting shape public understanding of projects by making information about a project readily available. The process of assembling content for a website contributes to a project's self-awareness. In negotiating among themselves about the materials needed to represent the project and to provide the context within which data are collected and interpreted, materials for online presentation become a scholarly product, a pre-archive project collection. The following examples underscore websites as an element of digital infrastructure for contemporary research projects. They serve as information environments, pre-archive activities, communication forums, and sites of catalog formation and delivery.

A website as a project information environment

Project websites represent a contemporary mode of data and information delivery. In the past, field projects often would end with the writing of reports, papers, and a new proposal, in addition to remaining as memories of lived experience in the minds of researchers. With the advent of the World Wide Web, it became traditional for LTER network member sites to develop a project-specific website. Each site made arrangements independently for local website hosting.

They developed processes for preparing and posting web content. The website provided participants with easy access to project materials and before long to data. The web was a welcome alternative to the technical overhead that accompanied user accounts and passwords for login to a local project system.

The EcoPrairie website developed into an integrative project forum with text and visuals. The data manager coordinating the web content was close to where data was generated and used by project researchers. Presence in the everyday activities of the project meant this intermediary heard about and addressed local technology and content concerns as well as witnessed many of the ramifications of in-situ data arrangements. The website presented five categories of materials: project activities, research, data, publications, and education/outreach. In addition, EcoPrairie adopted the LTER practice of extreme transparency by posting project proposals, interim reports, and some site review materials. Content grew as members of the project processed, interpreted, and assimilated materials. The website content was expanded and structured to display the products of the discursive process. Updates were in sync with changes in project priorities and insights. As a project information environment, discussions about the website fostered holistic reviews and understandings of the project. Collective internal review of how to convey project insights grew to include awareness of how is was used for project evaluation in external reviews.

A website as a pre-archive activity

A local website takes on added significance when viewed as a project collection, that is, an assembly of materials that captures and represents the project as a whole. What began during the EcoPrairie closing period as the digitization of old reports and photos, evolved into a realization that many project artifacts bundled together as digital objects together with datasets could be packaged into a highly-structured project collection. The general notion of 'data care' (Karasti et al., 2006) was extended to 'project collection care' with a collection recognized as adding value to datasets by capturing the scientific context:

> I'm thinking about the collection in terms of providing the context for the data. I started thinking about the series all together. So series in the collection – the photographs, the data reports, the species lists - emerged from what we found when we looked around on our servers and in our own drawers. (DM)

Ready access to a project collection was deemed valuable for subsequent use of the data in providing insights into the data and its interpretation. While the EcoPrairie information management team was required to upload datasets to the LTER Network Information System, preserving the project collection became a self-defined task of increasing importance as the project end became a settled reality.

The question of preserving the EcoPrairie website was elusive during the closing period since there was a lack of experience and policy for website preservation by local and campus IT units. Instead of website preservation, much of the website content became digital objects in the project collection preservation effort. Regardless of how website materials are captured, a long-term dimension is added to information environments when a website is planned with preservation of a project collection in mind. When there is sensitivity to 'project collection care' during the course of a project, the website becomes what may be considered a pre-archive activity that contributes to collection formation.

Website as a communication forum: primary and final websites

For projects in networks, websites serve as communication forums but there are issues of website control that arise with regard to local and network design. On one hand, websites under local control capture project views, activities, and accomplishments typically with flexibility and rapid updates. Local control also involves the responsibility of identifying and supporting web expertise. On the other hand, a centralized set of network managed member websites can be supported by a single Web master as well as standardized for ease of browsing.

EcoPrairie like all members of the Long Term Ecological Research (LTER) Network, maintained a website that was an integral part of its work and included in periodic evaluations (LTER IM, 2009). For members of the LTER Network, websites were locally controlled so that project staff were closely connected with the presentation of content. The capacity to represent the 'story' of a site grew over time, often driven by ongoing scientific activities and concerns. A website served as a project communication tool for a loosely coordinated set of project participants and sometimes served as a record of the site's history. By responding to local priorities and interests, local control introduced a degree of diversity across websites. The member sites within another network provide a contrasting example. The LTAR Network website is managed centrally by the government agency that coordinates the sites. Their website

is relatively new; it presents a list of funded projects, with each project linked to their proposal to become a USDA/ARS member site. It remains to be seen how the standardized web page format currently hosted at a centralized location develops.

For AtmChem, two online web representations had differing intents and overseers: a project science website developed by researchers at AtmChem and a project landing page developed by the external archive staff at AtmDM. The project science website was an information environment, supporting coordination and communication for an ongoing project. From the standpoint of project researchers, it was the primary project website and under their control. A project Webmaster selected by the research team taking into consideration project resources as well as a candidate's ties to project scientists, ensured the website evolved in concert with project activities. For the project landing page at the external archive, AtmDM staff were in control. They gathered together all the project-related materials as a whole collection and parsed some of the information into a standardized format. Standardization facilitated cross-project work in terms of both a human's ability to browse many projects and a machine's ability to access well-structured information. From the AtmDM perspective, the project landing page was a 'final' project web page that provided completeness with a list of 'related links' to many websites including the 'primary' project science website, data website, education website, and facilities website.

Researchers generally did not recognize the distinction between 'primary' and 'final' websites. In making the 'science web page' one entry in a flat list on the final project landing page, they perceived a loss of the primacy of the scientific driver that spawned related efforts. Despite the benefits of standardization, some scientists' considered the scientific work superseded in a manner that seemed to usurp the identity, status, and power of science. From this perspective, AtmDM as the archive preserving the project data, was perceived at times as an intruder in research efforts. Researchers were still assimilating the differences in intent between the primary and final web presentations. They did not find the distinctions intuitive between the need for an actively changing versus a preserved snapshot or for customized design versus a standardized template.

Website for delivery of catalogs: local and centralized

Digital catalogs delivered on websites by both AtmChem and AtmDM, highlighted the need for archives with control of metadata and data to take into account services from the perspective of data providers. AtmDM maintained an organization-wide master list of the field projects that it had supported including those of AtmChem, other AtmCenter units, and the greater atmospheric science community. The catalog permitted filtering based on keywords, temporal extent, spatial extent, and subjects to facilitate discovery and reuse of the data. From a data provider's perspective, however, filtering by organizational unit was not available. Its absence created a burden for an organizational unit such as AtmChem that had to create an independent catalog of their own projects.

At the beginning of this study, AtmChem presented its catalog of field campaigns in a calendar-like visual layout. By mid 2014 a move was made to a simpler, flat list with one line or record for each project, similar to the AtmDM project catalog (AtmDMPC, nd). The AtmChem web page was a catalog of projects with each record containing five columns (AtmChemPC, nd). This new catalog revealed significant vocabulary development including adoption of some of the AtmDM conventions. A 'Data Archive' column made visible a complete list of external 'archives' with which AtmChem partnered. This provided visibility and a standardized vocabulary for the names of the archives with which the laboratory had relations such as AtmDM, NASA, NOAA, CDIAC, and AtmChem.

The online catalog documents and makes visible the diversity of AtmChem data archive arrangements. In addition to prompting project researcher's to consider their set of projects as a whole, the focus on delivery of information to a local website required project participants to develop new language in considering collectively how to organize their project information. The entry 'AtmChem' in the archive column made visible the occasions when a project's data was held locally though this referred to local data storage rather than a managed data repository. The designation of 'AtmChem Archive' was a living reminder of the need for future discussions about AtmChem as a temporary 'stopping point' for data as expressed by the systems group (e.g. a data storage service) versus a 'final stop' as perceived by some researchers (e.g. an archive). In addition to prompting project researchers' to consider their fieldwork activities as a whole, the focus on delivery of information to a local website required project participants to develop new language in considering collectively how to organize their project information.

5.2.5 Internal relations: Science-data work models

A fifth basic element of data infrastructure is the relation established between science-oriented workers and data-oriented workers in the Local component. The four panels in Figure 5.1 model science and data work internal relations observed in the local component of the three cases. Data analysis and data production are shown as two distinct kinds of project data work. The figure over-simplifies data work in the local component. Data analysis indicates a kind of work that contributes to knowledge production in panel a. Incorporated in the term 'analysis' is a loosely coordinated set of activities central to research that includes selecting data for analysis, cleaning, checking, and formatting data as part of knowledge making. A typical outcome is the production of knowledge that may be published as a scientific peer-reviewed product such as a paper.

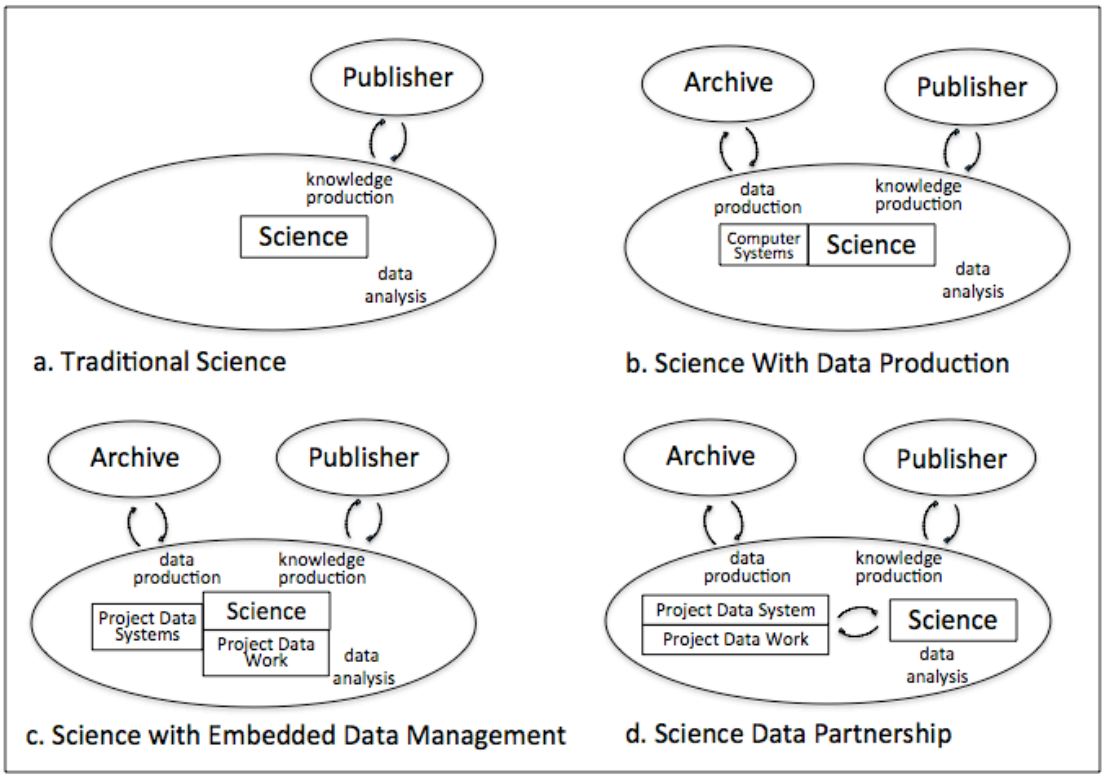

Figure 5.1: Four models of science-data work relations within the local component (large oval) for the cases studied: a. Traditional Science (EcoRiver); b. Science with Data Production (AtmChem); c. Science with Embedded Data Management (EcoPrairie); and d. Science-Data Partnership. Relations to organizations external to the local setting, namely publishers and archives, are also shown.

External to the local component, dissemination of papers often occurs via publishers that coordinate peer review, editing, and journal access. Data production, portrayed as having increasing complexity in panels b-d, is shown associated specifically with preparation of data for sharing, in this case with a data archive. Individual researchers in panel b carry out data production while in panel c is shown as a collective activity involving assembly and coordination of project-related data so is labeled 'Project Data Work'. In panels c and d, a project data system supporting project data work is overseen by data professionals focused largely on project data production. In panel d, the multiple project systems often are more tightly coordinated so are depicted as a 'Project Data System'.

Traditional Science model

For EcoRiver, internal science-data work relations followed a Traditional Science model. An assistant, delegated the job of assembling data into text files or excel spreadsheets, supported a scientist conducting research. One research assistant, who was a field participant and a laboratory technician, described the EcoRiver work as task-oriented and somewhat open-ended but the work was not considered data management:

> My primary job duties are sampling, managing samples, and handling the data, that's the majority of my work. Although I branch out in other areas as needed. … I'm a data recorder and organizer. I just don't think of it as data management. I just think of it as, I put the data in the sheets and put it in an order that I can find it. To me I think a data manager is working with data that's more different. Everything that I'm working with, it is very, very similar. And I'm thinking if you are working with different projects, from different people, you are getting more, more input from different sources. (PS)

In the EcoRiver case, sampling procedures were documented and stored separately from data on a university, restricted access computer system. The final data product was a workbook set of spreadsheets that resided on a laboratory machine containing all the data, analysis, and derivations. The researcher typically kept a copy of the workbook with which to work. The assistant described the workflow:

> Well I usually put all the data into one spreadsheet, just to keep track, so I know where it all is and so I have pages and pages and pages ... And then when somebody needs something, I pull a couple of pages out of it [the workbook] for them. (PS)

The majority of the data, though backed up, did not reach an external archive.

The first panel Figure 5.1a presents a generalized model of Traditional Science where data is available for local use. Data activities are integrated into the research workflow within a project, laboratory or department. This model presents a science-driven or curiosity-driven research effort with a principle investigator overseeing and taking the lead in most aspects of the research including work with and use of data. Outcomes are scientific publications. Data is not disseminated but shared on request.

Science with Data Production model

AtmChem internal relations included science and data production. One researcher at AtmChem referred to work with data as following a 'basic model': "We put the instrument on the plane, we go take the measurements, bring it back, do the data processing, and then the analysis, and the science, and publications after that." The basic model, however, in practice is an 'extended' model that includes data production. There is no project-designated data manager so researchers depend on assistance from junior scientists as well as archive staff for submission of data to an external archive. One archive staff member commented on the relations between those at the archive and those in the project:

> It's kind of like data management is its own little entity, and science is its own entity. And they have to work hand in hand. (SE)

The second panel Figure 5.1b presents a generalized Science with Data Production model. It represents a science-driven research mode of operation involving data analysis. The use of data results in the production of knowledge and publication of findings as with the Traditional Science model. The addition of data production brings new activities including selecting data for archive, packaging, and submitting data to an archive. Researchers engage with the archive individually. Ties exist to one or more external archives.

Science with Embedded Data Management model

The EcoPrairie site was configured with an embedded data manager. Data management began in the early years (1982-2006) as a part-time task for an associated research scientist and then was assigned to a staff research assistant. In later years (2006-close), as described in this

section, it became a full-time position that included supporting data coordination for researchers, students, and volunteers. A resource manager associated with EcoPrairie explained how the project-related data management position changed over time:

> First we had part time people and for a while the data manager was also a researcher. The milestone is when we had a dedicated person. There was even more continuity when she went full time, and left the fieldwork, and just did data management totally. (RM)

A resource manager referred to the full time commitment of a data manager as a 'big change': "[The data manager] got involved with the network and all the other sites and brainstorming about new systems to use to store data, keep up data, maintain data."

The EcoPrairie data manager remarked on understanding the data role as possessing professional responsibilities distinct from fieldwork and data analysis. Activities associated with collective data work that were mentioned by members of the information management team included 'working with the scientists in the field and in their offices' and asking repeatedly "where are the data sheets?" until they are in the hands of data personnel. Additional tasks involved helping with GPS equipment, supporting computer and software needs, developing and maintaining field crew notebooks of procedures, overseeing data entry by students, formatting and documenting data, cleaning data, and creating variable definitions for measurements. Project researchers at EcoPrairie described situations where data services were needed that did not exist so a portion of the data work involved planning, design, development, enactment, and documentation of new data procedures and capabilities. Throughout these activities, proximity and everyday interaction with local researchers facilitated discussions on a regular if not daily basis with those who would be using the services.

The third panel Figure 5.1c shows a local component collective data arrangement described as a Science with Embedded Data Management model where data work is identified as a separate task and delegated to an individual other than the researcher. The role is an integral part of research activities. The formalization of a data position is indicated by enclosing 'Data Work' in a box as a separate category of work. The data manager is at the center of collective data activities and growth of data infrastructure. The development of new data practices and attention to data workflows facilitates data assembly, eventually becoming identifiable as a data system shown within a box as it develops and formalizes over time. While the use of data for local scientific research that results in publication of scientific findings remains a high priority,

data production is a process managed by a data intermediary who handles interactions with archive partners on behalf of project researchers. The data manager may initially be involved in fieldwork, instrument work, and data analysis but becomes increasingly responsible for a growing set of data management and data system tasks associated with data from multiple sources. There is generally a collegial but hierarchical relationship where science leads and data management occupies a support role. For this model, data management typically does not have an independent operating budget within the organizational unit.

Science-Data Work Partnership model

During the termination period at EcoPrairie, the hierarchical relationship of scientists and data professionals evolved into a flattened partnership. In maintaining everyday ties with local researchers, the information manager created and tended a data workflow from researchers to a project-oriented data system. Rather than individual researchers interacting with archive staff, the data system is aligned with at least one remote archive for data submission. At project close, one researcher explained data management as follows:

> When the site shut down, when we were going through that whole process, I think it was reassuring to know that one of the areas that we still did really well on was the data management side. … when we have all these other issues that are going on. I know that with the shut down going on for many it was pretty tough to have the site be gone but I think that at least we knew that data management had that part of the ship pointed in the right direction and that never flagged. (RS)

With the closing in sight, researchers turned their attention to participating in and developing other projects. At this point, the EcoPrairie information manager gained some independence as well as increased responsibility in addressing data issues and was invited to join as a co-investigator on the final project grant.

The fourth panel Figure 5.1d presents a generalized Science Data Partnership model. The boxes indicate formal recognition of data work that includes a data system. The significant change in scope and status of data work from partnering of science work and data work is illustrated by putting data work at the same level as science work. Their ongoing interaction is highlighted via a set of bidirectional arrows. In this model data management may benefit from a separate budget dedicated to data work with signatory authority. Following the notion that

delegating authority in addition to tasks raises up leaders, a data-related budget spurs data work and innovation.

5.2.6 External relations with archives

Having explored the internal relations between scientific research and data work, the sixth element characterizing data infrastructure for collective data work addresses the larger context created by the relations of a project's local component with external archives. The panels in Figure 5.2 show salient features of the local component in addition to the connections to external data archives. Three shaded ovals on the left are place markers that represent the fieldwork of each case. The number of data-related intermediary roles are indicated by small ovals that included data managers, Web masters, technical staff, software engineers, research scientists, and data curators. The intermediaries have a variety of functions ranging from work with technical systems and web delivery to data assembly and documentation. The dashed lines to archives indicate submission of data to external archives by individual researchers while solid lines indicate a project data intermediary working with a project data system making submissions.

Differences in the project's local collective data work and relations with external archives are evident. The dashed lines for AtmChem indicate individual project researchers make submissions to archives. Its local component supports researchers to work on data in individual or in ad hoc groups but requires researchers to submit their own data directly to the various archives. For EcoPrairie, the data manager oversees submission of datasets from the project data system to archives.

EcoRiver: Nascent local infrastructure and no relations with external archives

The first panel Figure 5.2a portrays the nascent data infrastructure of EcoRiver, a project without a community-wide data assembly process in place. The project website was hosted by a museum partner. The formalization of collective data work began as a required part of data management planning associated with EcoRiver Field Station and with a two-year research award at EcoRiver. Groups at this location did not have shared technical infrastructure available for collective data assembly. A university-wide enterprise license for file sharing software (Box,

nd) provided group access and data storage that facilitated assembly of data isolated on individual computers.

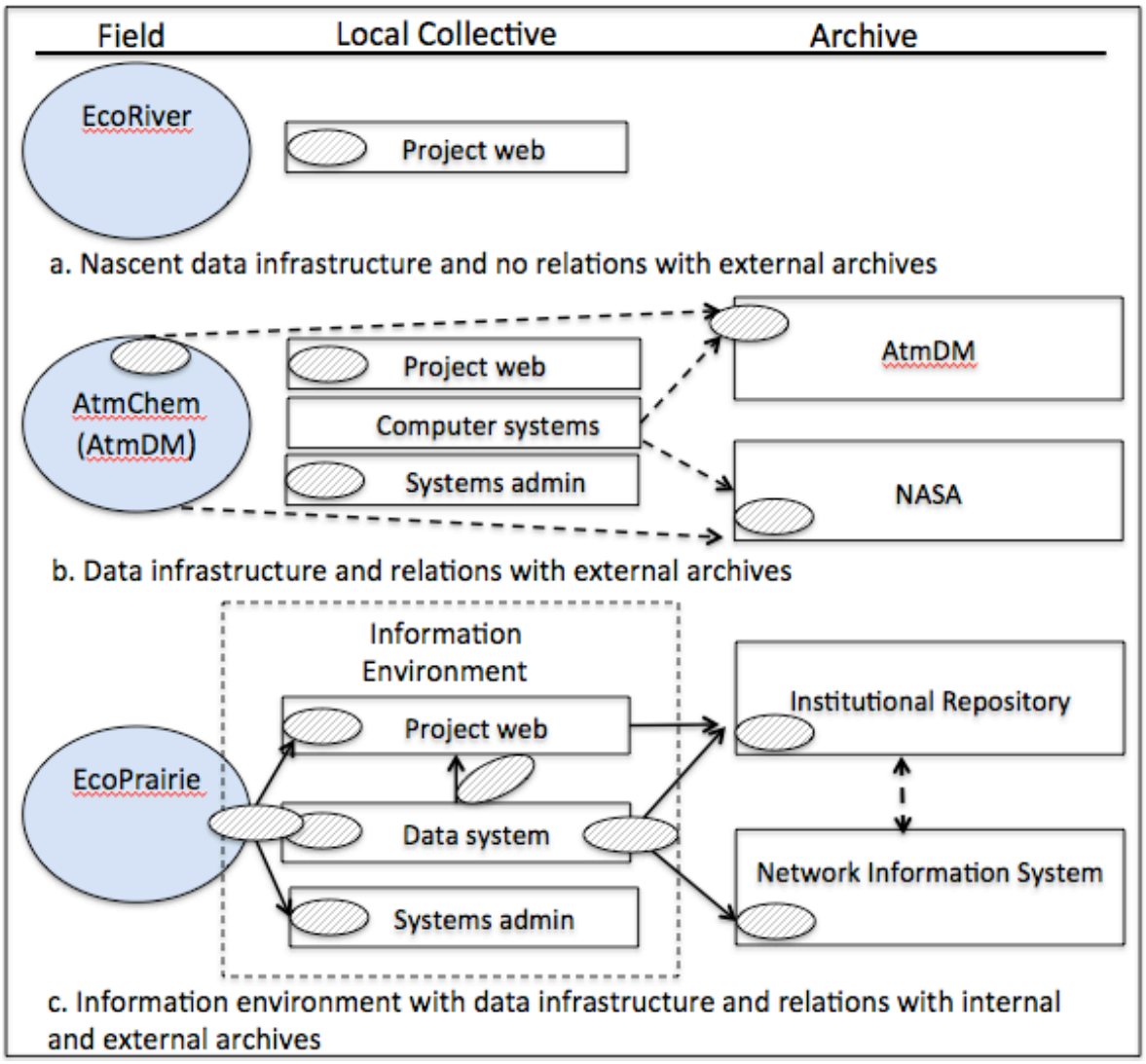

Figure 5.2: Data work configurations for the three cases are shown in separate panels using the field-local-archive organizing framework. The names of the case appear in the field component. The dashed lines to archives indicate individual project researchers submitting data to archives. The solid lines to archives indicate submission of data to external archives by a project data intermediary working with a project data system. Small ovals represent data intermediaries.

An issue evident at EcoRiver was the lack of contemporary collective data assembly for data sharing both at the preserve level as well as at the organizational level of the individual researchers. Members considered collective data work at EcoRiver as separate from data management at their home institution. In some sense, the preserve office served as a network

office that coordinated a set of project members. Given these circumstances, it was not possible to discern whether collective data work would continue to develop in sync with or at the same pace for network and member levels.

AtmChem: Local infrastructure and relations with external archives

The second panel Figure 5.2b shows the data work configuration for AtmChem. When AtmDM worked with AtmChem, they supported data activities in the field and also were available to provide archive services. With aircraft used for fieldwork, both science project instrument teams as well as a software engineering team were involved in managing data and technical issues throughout the project. Data sharing within the group began in the field. This enabled assessment of instrument performance and data quality. Three versions of datasets were transferred to the archive successively: field, preliminary, and final datasets. The data generators prepared their datasets for submission to a project-designated archive. Within the Local Component for AtmChem, two elements of data infrastructure are shown: web support and systems administration. The website for AtmChem had a local Web master for their website.

There were a number of external archives with which AtmChem coordinated, two of which are shown by way of example. Data intermediaries are shown: two within the local component and one within the archive components. An AtmDM software engineer, available in the field and post-field, is designated as a data intermediary for a particular research project. In an alternative archive arrangement, a NASA research scientist served as data contact for the NASA facility. In this configuration, the external relations are portrayed as dashed lines that reach from the individual project researchers to external archives though in some of the smaller campaigns, as mentioned earlier, AtmChem uses its local data storage.

EcoPrairie: Local collective data work and relations with external archives

The third panel Figure 5.2c illustrates the data work configuration of EcoPrairie where four aspects of data infrastructure support the local component are shown: data management intermediaries, systems technical support, a data system, and the project web pages. The coordination and information flow supported by multiple intermediaries is noted as an information environment by a dashed rectangle enclosing the interdependent ensemble of data

infrastructure elements. The work of a number of intermediaries involves communication and interactions across project, laboratory, library, and university units.

The data manager, familiar with project fieldwork, local data systems, GIS activities, and the website, played a major role in creating the information environment as well as ensuring the movement of data from the field to the local data system. Individual investigators and field volunteers sent data files to the data manager who gathered and checked data files, created metadata from field logs, designed and updated crew manuals, and held discussions with those working in the field. A single, multi-function data system was managed by the data manager who frequently checked on project and community needs as well as department, university, and network technical services available with an eye toward planning forward for redesign, development, and maintenance of project-related data infrastructure elements. The data manager interacted not only with researchers but with the other contacts shown in the local component where support for technical infrastructure was provided by the local university unit, the university-wide IT unit, and the LTER Network. Both prior and subsequent to the development of the network information system, data was delivered locally via a data catalog on the project website together with an assembly of project materials.

Two external archives with different approaches to preservation are shown as recipients of EcoPrairie data. There are intermediaries at the archives supporting submission of data and metadata: a data curator at the institutional repository and technical staff at the LTER NIS. As the LTER Network Information System (NIS) developed, the local data system was tailored to deliver project data to the NIS. In terms of the data ecosystem, it is important to note that the development of the NIS did not replace the local data system. Rather, community data work in close proximity to project participant activities continued and expanded to include managing the process of submission to an archive. Changes in scientific data needs were identified and addressed locally while meeting external requirements at archives that were handling larger quantities and greater diversity of data from many projects.

In the quest to understand the complexity of a natural system, EcoPrairie researchers collaborated with a mix of individuals associated with other projects, networks, and archives. While Figure 5.2c shows EcoPrairie data moving to the two major archive partners, the EcoPrairie information management team and individual researchers provided data to other data facilities as well. Figure 5.3 provides a more complete picture of EcoPrairie relations with

external archives where data movement is grouped into three clusters: A. Data files of researchers for personal work; B. Collectively managed datasets via the project data system and sent to project partners; C. Individual sharing of data by project researchers with other networks.

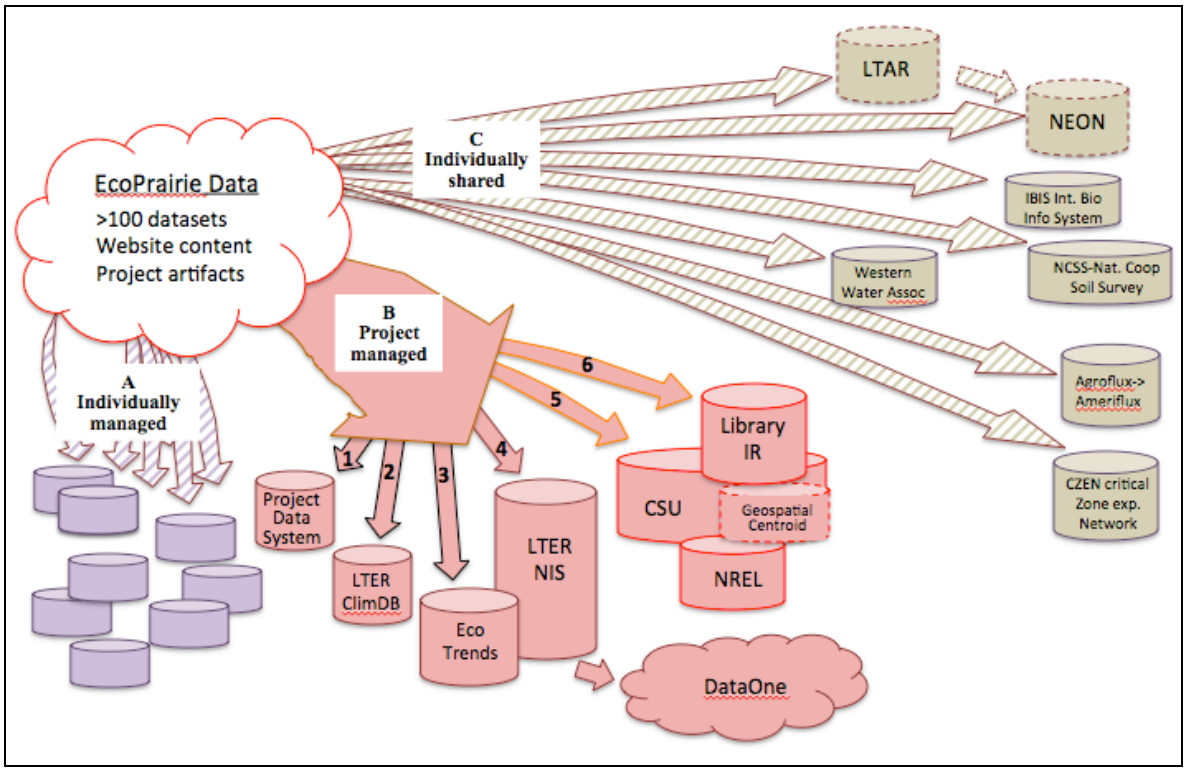

Figure 5.3: The many pathways of EcoPrairie data are shown in three clusters: A. Individually managed for personal use, B. Project managed and shared with partners, and C. Individually shared by researchers with alternative networks. The hashed arrows depict relations carried out at the discretion of individuals. Dashed cylinders indicate systems in development.

The three groups are described in the following list:

A. Individually managed by individual researchers for personal use
    A1. Datasets to individual storage devices
B. Project managed data that is sent to project partners
    B1. Datasets and project artifacts to EcoPrairie data system and website
    B2. Datasets to the LTER specific climate data system (Henshaw et al., 2006)
    B3. Datasets to the LTER/USDA EcoTrends data system (Peters et al., 2013)
    B4. Datasets to the LTER Network Information System as a node on the DataOne network of networks (Servilla et al, 2016; Michener et al. 2011)
    B5. Datasets transformed to a non-dynamic form and stored on project storage unit; the Geospatial Centroid (nd).

B6. Project collection of datasets and artifacts to the Institutional Repository (Kaplan et al., 2014b)

C. Individual sharing by project researchers of datasets with other networks
   C1. Critical Zone Experimental Network (CZEN)
   C2. Ameriflux Management Program (Ameriflux)
   C3. Western Water Association (WWA)
   C4. Nutrient Cooperative Soil Survey (NCSS)
   C5. International Biological Information System (IBIS)
   C6. National Ecological Observatory Network (NEON)
   C7. Long-Term Agro-ecosystem Research (LTAR) Network

This broader view demonstrates that in addition to closely managed relationships with one or two archives, data is migrated to many other locations. With the same data residing in multiple locations, tracking the data is difficult at this time although ongoing work with digital object identifiers may change this situation significantly.

5.2.7 Dataset and project collections

A seventh element of data infrastructure for the local collective data work involves data collections. Two distinct categories of archives based on their content were distinguished: dataset collection and project collection. Archive policies designate what kind of digital objects the archives will accept for preservation. Dataset collection archives accepted submission of datasets (data files and related metadata using a designated standard). Project collection archives accepted datasets as well as artifacts such as photographs, field notebooks, species lists, geospatial maps, reports, brochures, meeting programs, and posters, all in various file formats. The dataset collection narrower focus on datasets in a common format facilitates development of advanced services such as discovery and query of datasets by variable. In contrast, a project collection incorporates a variety of kinds of artifacts and makes them available for online browsing. Examples of the two categories are given for archives used frequently by EcoPrairie (Figure 5.2c) and AtmChem (Figure 5.2b).

Dataset collections:

- EcoPrairie and the NIS archive

  EcoPrairie submitted datasets to the highly structured LTER NIS system that included search capabilities. A metadata validator with reports on whether data met the community

criteria included a link to data files. The metadata validator helped submitters prepare high quality documentation. This set the stage for query by variable, a system feature that had not yet been implemented. EcoPrairie metadata delivery to the LTER NIS was enabled by the project data system capability of exporting EML, the metadata standard used by LTER NIS.

- AtmChem and the NASA archive

  AtmChem project researchers submitted datasets to the NASA archive using an upload process that included a format check of the self-describing data files that were required to conformed to the **I**nternational **C**onsortium for **A**tmospheric **R**esearch on **T**ransport and **T**ransformation (ICARTT) file format. The datasets were displayed with information about the individual investigator, variable names, and research description available for browsing. The data files were downloadable together with data products such as merge files created by the NASA archive.

Project collections:

- EcoPrairie and the Institutional Repository

  EcoPrairie submitted datasets and project artifacts to the university library that was using a collection management application for its institutional repository. They were exploring development of a data repository though online ingestion of digital objects was not available. Query of the data by variable was not part of the planning with limitations including the software as well as the variety of kinds of items accepted and their differing formats.

- AtmChem and the AtmDM Archive

  AtmChem submitted self-describing datasets in formats such as the ICARTT for chemistry data as well as HDF for larger volumes of satellite and model data. Due to the size of holdings and the differing formats accepted by the AtmDM archive, data was not available for online query by variables. In creating project landing pages, a variety of information about the project and its related links was assembled for browsing.

While EcoPrairie followed the LTER community convention of submitting datasets to the LTER NIS, the closing period included migrating project data to an institutional repository as a project

collection (Figure 5.2c). In making this move to expand beyond the LTER cultural norm, there was a trade-off made in terms of time spent with each approach. In contrast, AtmChem project researchers had a history of working with research data archives. The decision as to which archive to work with for any one project was influenced often by field logistics support and resources.

5.2.8 The role of intermediaries and infrastructural thinking

The eighth element of the data infrastructure for the local collective is the role of the data intermediary. While roles associated with data work are evolving, academic programs are developing (NRC, 2015). In universities there are new degrees in areas such as data curation and data science. In discussing the concept of an information community, Durrance et al. (2006) explored the role of information intermediaries while Mayernik (2016) identified intermediaries as an institutional carrier, that is, one of "the entities (human, nonhuman, and conceptual) that create, constitute, and perpetuate institutional systems". In practice, the configurations portrayed in Figure 5.2 include some data intermediaries that are half in and half out of a component. These are a visual reminder of the data-related intermediaries carrying out boundary crossing work that supports the movement of data.

Assembly of data for a project requires recognized leadership and a shared vision. In the three cases for this study, part-time or full-time intermediaries, including data managers, data facility engineers, and research scientists, provided leadership for collective data work. All demonstrated infrastructural thinking, that is, planning forward for data work by taking into account evolution of data services, technologies, webs of repositories, and networks of collaboration in the larger data landscape. They designed short-term strategies nested within a longer temporal orientation, a circumstance described as 'infrastructure time' (Karasti et al., 2010).

At EcoPrairie, represented by the Science with Embedded Data Management model (Figure 5.1c), a project data manager looked beyond program-specific data planning at site closing to explain the time spent in a trade-off of two activities, a two-year deadline and a longer-term strategy:

> It's not two phases but two different types of activities…One being decommissioning and the second a launching out from that as an opportunity to learn about building up an infrastructure to support data management and curation for the lab. (DM)

With this insight, the two-year closing period became an opportunity both for bringing data management efforts to a close in a professional manner but also a time for inquiry into infrastructure options for meeting future, laboratory-wide data needs. The data manager had a proactive attitude: "My approach to decommissioning, I'm trying to make opportunities out of it." The response shows project termination taken as a challenge to begin a new kind of planning. The project termination expanded the information manager's concern for a single project and a single network to working with initiatives external to the project. A commitment to open up options for EcoPrairie data was evident:

> I'm trying to play in as many sandboxes as possible in hopes that something will work out, something will stick. There's no perfect solution. At the end I have to have it [the data] existing somewhere. I can't think of this as a trial. (DM)

Infrastructural thinking by this intermediary spurred work on both a project collection as well as the concept of interoperability. Project researchers mentioned that the majority of project artifacts would have ended in trash bins if the scope of EcoPrairie data efforts had not expanded during its closing period. The two-year supplement for closing afforded the time required for forward planning. While data submission to the LTER NIS remained a key responsibility and assessment criterion, EcoPrairie also developed an approach to addressing preservation of a project collection that included a more complete set of project artifacts to complement their collection of datasets. This in turn addressed concerns with sustainability of data facilities through partnering with a domain repository *and* an institutional repository. In carrying out data work within the broader data landscape, awareness of alignment between systems grew. Design thinking originally focused on data access provided by one archive opened up to considerations of distributions of data work across data work arenas when data holdings were staged for data system interoperability.

At EcoRiver, represented by a Traditional Science model of science-data work relations (Figure 5.1.a), researchers and managers demonstrated infrastructural thinking in the absence of designated data intermediaries. The notion of collective data management was discussed with

partners who already shared a strong commitment to ecological stewardship of the preserve that involved managing a complex, human-natural coupled system guided by scientific research. Several science leaders recognized data stewardship as a timely, synergistic concept for their collaborative work and the increasing amounts and kinds of data at the preserve. They included data management in planning for a new field station, annual science meetings, a data stewardship workshop, and a five-year strategic plan for EcoRiverOrg. Despite the varying states of organizational infrastructure supporting the various partners, first steps toward a vision of collective data management were taken.

At AtmChem, with the Science with the Data Production model for internal science-data relations (Figure 5.1b), there was concern with organization of a center able to raise awareness of the critical collaborative nature of their scientific activities. They encountered difficulties in conveying the blend of integrative research stretching from instrument building to field as well as laboratory measurements, from data use in atmospheric chemistry studies to data use in modeling, and from development of models for understanding specific field arenas to development of models that advance understanding of the atmosphere more generally. Initial plans for the center included consideration of data arrangements in response to an ongoing review, but the majority of participant time during my fieldwork period was dedicated to arrangements for support of the atmospheric science community across many academic institutions. Though AtmDM identified this period as a potential opportunity for dialogue with AtmChem about data activities and initial interactions occurred, the individuals and timing involved were not opportune for joint data discussions. AtmDM attention to intra-organizational relations was a recognition of a need to improve relations with project researchers, that is, to bridge the gap between research centered and data centered organizational units discussed earlier.

5.3 Findings

Drawing on cross-case analysis and using an analytic framework with three components, eight elements of data infrastructure have been identified that characterize local collective data efforts (Table 5.3). The summary of the elements of data infrastructure for the local component

are followed by an overview of the elements particularities for the data infrastructure of the three cases (Table 5.4).

5.3.1 Data work configurations

Bringing together the three cases provides a window into the variety of data work configurations found in practice in the field-based natural sciences (Figure 5.2). These configurations represent the context within which a local data infrastructure develops. EcoRiver, an ecological partnership with a growing need to assemble data, illustrates initial data management efforts in coordination with multi-sector partners involved in regional and global river networks. AtmChem, an organizational unit that participates in a variety of field campaigns each year in targeted locations, worked with a number of external archives. EcoPrairie's highly developed local data management included a project data system.

Data landscapes dominated by "macro-constructions of a larger social order" (Marcus, 1995) are in a position to overshadow design of local collective or mesoscale data infrastructure. In taking account of all three components, the analytic framework sets up one of the tricks of the trade discussed by Star (1999) and Suchman (1987): to make work visible that is largely invisible. In particular, the field-local-archive framework is inclusive of contributions by all those involved in data production and by the project data infrastructure. The local component is preceded by a field component with a field team that generates data; it is followed by work at remote data facilities that preserve and provide access to data from many sources. From the project perspective, the framework places the local component within a larger context that takes into account where data are generated as well as where data can be preserved. The project's work in the local component, functioning as both post-field from a sampling perspective and pre-archive from a preservation perspective, represents a way station that bridges the field and archive. From a field perspective, the mesoscale setting is a destination for continuing review and analysis when fieldwork is finished and sets of data files generated are complete. From a preservation perspective, the data work carried out in the field and local components is critical to data product formulation and documentation.

5.3.2 Elements of data infrastructure in the local component

The elements of data infrastructure identified during this analysis and elaborated in the previous sections are summarized in Table 5.3. The focus on infrastructure relating to data work in particular highlights categories that differ from but overlap categories discussed by Star and Ruhleder (1996) presented in Table 5.2.

Table 5.3: Elements of data infrastructure supporting the local component

#	Elements of Data Infrastructure
1	Data management
2	Systems management
3	Data assembly and data systems
4	Web site and information environment
5	Internal relations of science & data work
6	External relations with archives
7	Data collections
8	Data intermediary roles

An overview of the observations of the elements of data infrastructure that occurs within the local component for each of the cases is presented in Table 5.4. The first four elements are assessed as developing or developed where developed is further described as carried out by project researchers individually or with archive staff. The final four elements describe differences across the cases in terms of internal and external relations, kinds of data collections, and the roles of intermediaries. Comment on the completeness of the list would require further case analyses.

Table 5.4 is a step toward understanding the different states and the varied approaches possible for data work configurations in the field-based sciences. EcoPrairie can be described as having developed local data infrastructure. EcoRiver is shown with two elements of infrastructure that are developing. AtmChem, on the other hand, is shown as developed. In this case, however, the 'local' component of data infrastructure for the project is supported both by participants affiliated with the project but also by participants from the archive. Their archive partner reaches beyond the category of archive component to provide services in the field as well as in local arenas. The table shows the blurring of boundaries that occurs in the AtmChem local component where the project data collective is supported by both local and archive components.

Table 5.4: Elements of data infrastructure in the local component for the three cases

#	Elements of Data Infrastructure	EcoPrairie	EcoRiver	AtmChem
1	Data management	developed (local)	developing	developed (archive)
2	Systems management	developed (local)	developing	developed (local+archive)
3	Data assembly and data systems	developed (local)	-	developed (local+archive)
4	Web site and information environment	developed (local)	-	developed (local+archive)
5	Internal relations of science and data work	• science with embedded data management	• traditional science	• science with data production
6	External relations with archives	collective submits	-	individual submits
7	Data collections	• dataset • project	-	• dataset • project • multi-project
8	Data intermediary roles	• data management • systems admin	• research assistant • technical assistant	• systems admin

Although the three-component framework accommodates the field to archive movement of data for EcoPrairie, the category boundaries break down for AtmChem. This together with the dynamics and feedbacks of the system will be discussed in the next chapter where a data work system model is presented, an approach that further captures the complexity of data work configurations.

CHAPTER 6. A DATA WORK SYSTEM MODEL

A Data Work System model is developed in order to address the non-linearity and the changing nature of collective data work in research environments. A number of data work arenas are introduced that capture some of the complexity of project data work. The model features feedback mechanisms and the dynamics of interdependent systems of subsystems. The model encapsulates the context within which the local component resides. The chapter concludes with findings including trade-offs observed in infrastructure development, data gateways as mesoscale infrastructure elements, and three distinct kinds of data collectives.

6.1 A System of Data Work Arenas

Adopting a systems lens is critical in order to account for data work that is not only complex but also constantly responding to changing circumstances and insights. A Data Work System model is developed that recognizes each of the framework components (field, local, archive) is comprised of one or more data work arenas. Each arena is a site of engagement with project data. Each arena has boundaries or borders that define a segment of specialized data work carried out by a group of individuals familiar with each other as well as the arena's data tasks, routines, and history of events. A project consists of a patchwork of data work arenas that are coordinated formally and informally. In working with bench work laboratories of contemporary biology, Knorr-Cetina (1992) describes the link between work arenas as a border that defines internal and external activities. The system model makes clear the need to mediate at borders in order to facilitate the movement of data between arenas and components. To move data from an arena, articulation work contributes to coordination, that is, to alignment of activities across arenas so that data is able to 'flow'. Arena-to-arena interactions contribute to the dynamics often difficult to capture in data workflows for an entire field-to-archive system.

This study underscores ongoing data tasks, communication, and learning in considering a data work system as a dynamic ecosystem of interdependent data work arenas comprised of people engaged in work with data, technology, and other people. In discussing an 'ecology of mind' (Bateson (2000), an 'ecology of infrastructure' (Star and Ruleder (1994), and a 'data ecosystem' (Parsons et al., 2011), the notion of the unexpected or unaccounted for is added to

how views, situations, and systems unfold. The system in this study refers to dynamic elements subject to planned as well as unanticipated events, interactions, and configurations at various scales and scopes.

A systems approach represents an alternative that avoids some of the limitations of the lifecycle approach. Lifecycle models comprised of well-defined steps and tasks have been developed in order to envision work with data, create shared vocabulary, and provide graphic visualization of the processes involved (Higgins, 2012; Ball, 2012; Pryor, 2012; Carlson, 2014). Examples of augmented lifecycle models include adding a related data analysis cycle (Baker et al., 2009; Baker and Millerand; 2010), a scientific data processing lineage (Bose and Frew, 2005), a distribution of data management responsibility within scientific field research (Wallis, 2012), and the intertwined nature of research and data lifecycles (Tenopir et al., 2011). Models present an ideal representation so suffer from a focus on higher level organization with gaps between functional boxes (Humphrey, 2006). Many lifecycle models focus on higher level organization at archives rather than on project collectives closer to the data origin, that is, on pre-ingest or pre-archive arenas. Further, it is noteworthy that lifecycle models have proliferated as they have been tailored to individual archives (Carlson, 2014). At way stations that carry out pre-archive data work while embedded in science environments, the diversity of data arrangements with tailoring of data for local data needs has only begun to be documented.

6.2 A Dynamic Data Work System

The model shown in Figure 6.1, brings together multiple data work arenas identified during interviews, as interdependent subsystems. This Data Work System model is a representation of data work carried out in the primary case. The perspective is that of data-collecting projects with shared platforms for fieldwork. The model entails two major information environments (project and public) together with three components (field, local, archive) that encompass six distinct arenas (local planning, field, local postfield, local collective, archive, and alternate archives). The dashed circle shows the project information environment containing local and field data work. The solid circle indicates a public information environment where project-related material may appear in preliminary and published form. Though the focus in this study is on data work configurations rather than metadata in particular, metadata is key to data

discovery, integration, interoperability, and dissemination in information systems so is shown as a dotted backdrop (green) in all the arenas in order to convey the potential for contributions to be made in capturing metadata throughout the system.

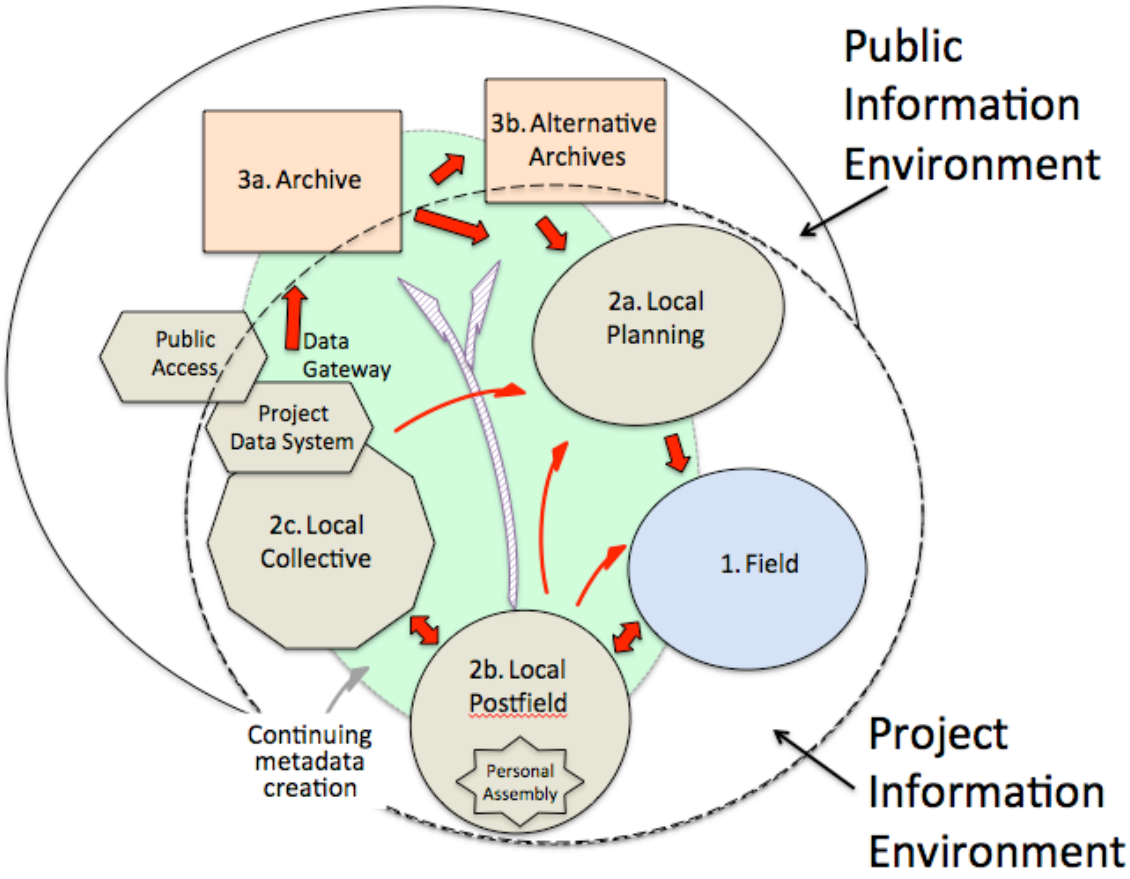

Figure 6.1: A Data Work System model for projects in observational field-based sciences is shown with both public and project information environments. Three groups of data work arenas are labeled: 1. field (blue), 2. local (brown), and 3. archive (orange).

Data arenas are nested within components. Arenas are numbered and color-coded by analytic framework's three-components: 1) field (blue), 2) local (brown), and 3) archive (orange). For the six data work arenas that make up the data workflow, increasing structure in data work arenas is indicated by the use of shapes: ovals for project activities prior to data assembly, polygons for arenas where collective activities are loosely structured, and squares for highly structured data work in archives. The bold red arrows indicate data movement inclusive of local collective data work. The thin red arrows represent informal sharing of data and the hatched

arrow shows an individual investigator's submission of data. The 'field' component is portrayed as a single arena where sampling occurs. The framework's 'local' component is defined by three arenas: 2a) planning, 2b) postfield, and 2c) collective. In this component activities are often carried out in geographic proximity to one another and to a science project office. Finally, the archive component consists of two kinds of arenas: a primary archive and alternative archives illustrating how data submitted to one archive may appear in another archive.

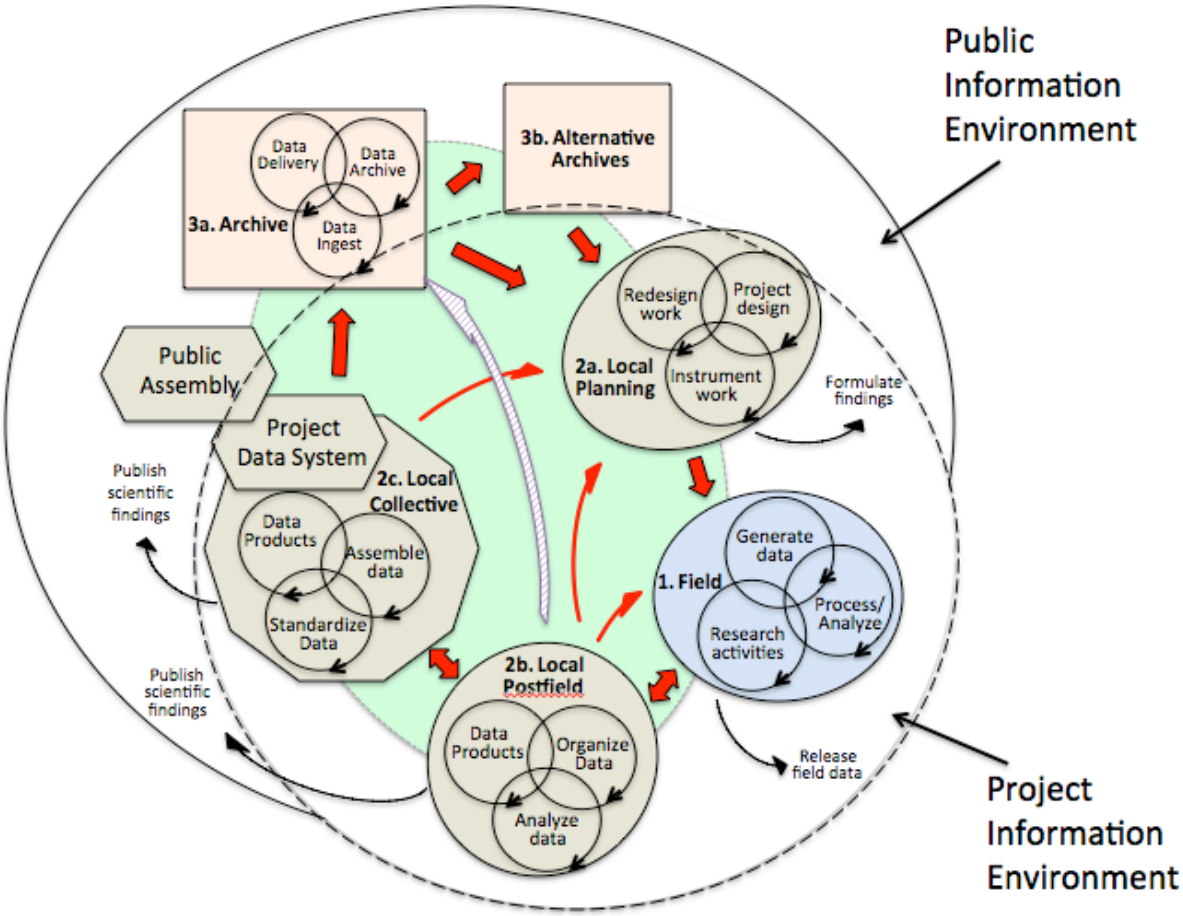

Figure 6.2: Data work arena dynamics are portrayed as subsystems within a project's data work system. Though oversimplified, three sub cycles are depicted within each data work arena to suggest the multiple interactive processes involved in working with data within arenas. Three groups of data work arenas are labeled: 1. field (blue), 2. local (brown), and 3. archive (orange).

Figure 6.2 adds dynamic detail to the Data Work System model. Each arena is shown as a subsystem of the whole that contributes to the project and public environments. Each arena is also shown with a highly simplified representation consisting of an internal set of interdependent

processes. A research project begins in the local planning arena (Figure 6.2-2a). Sampling design and instrument work play key roles in preparing for fieldwork. Periodically new circumstances, understandings, or requirements lead to redesign of instruments and sampling plans as well as to the composition of field teams prior to fieldwork. From here, researchers travel to the field.

The field arena (Figure 6.2-1) is the data origin where data are generated from observations and measurements. The unit of fieldwork bounded in time is typically described by a term specific to a discipline. EcoPrairie and EcoRiver used phrases such as 'field season sampling', 'time-series sampling', and 'pre and post flooding' while AtmChem used the term 'campaign' to describe a field visit. Fieldwork may be further divided into a set of related segments. For instance there were multiple flights within a single atmospheric field campaign. While in the field, data are often processed and analyzed in a preliminary or quick look manner. During field visits, project researchers may carry out collaborative activities – planned and unplanned – such as analyzing data together with other datasets, group discussions of data that elicit stories, and joint planning for the next segment of fieldwork. Teams with established field-laboratory communications may release data while in the field, bypassing the local postfield arena to deliver data directly from the field to venues closely coordinated with project activities. For instance, photographs may be sent to the local collective arena to be posted on a project website for education outreach or streamed sensor data may be posted directly on a public website.

Upon finishing a field campaign, project researchers return to local data work arenas. In the local postfield arena (Figure 6.2-2b) and the local collective arena (Figure 6.2-2c) researchers share datasets and stories. Project participants in these arenas, those who were in the field and those who were not, work with the data. As researchers who were not in the field interact with field participants and other project personnel, they become enmeshed in the data work and gain familiarity with the field campaign within the context of the project. Project groups often hold data workshops where data and findings are presented, discussed, and interpreted in support of integrative data work as well as development of new data products. Stories are told that convey individual insights into the data and experiences from the field.

In the local postfield arena (Figure 6.2-2b), work with data often involves cycles of calibrating, processing, cleaning, analyzing, and visualizing. Individual researchers may eventually publish papers or reports on their individual data findings. They may send their data

back to the local planning arena or the field arena to inform future fieldwork and future planning. In addition, they may send their data to an archive that differs from the project-designated archive. A bi-direction arrow shows datasets sent to a local collective arena since datasets may be updated as work with data proceeds for data generators and their colleagues.

When a project supports local assembly of data from many researchers, a local collective arena (Figure 6.2-2c) becomes a center of data activities. The infrastructure in this arena at minimum involves shared file storage and local data practices such as controlled vocabularies and format conventions. Responsibility for assembly and organization of project data is clearly delegated. As researchers and data specialists interact in this discursive, deliberative venue, updates and modifications are made to data over time. Data work may include inventories of data as well as identification, formatting, and selection of datasets to prepare for sharing the data. Some data is left behind. For example, in the local postfield, data may be left unanalyzed due to time constraints or found untrustworthy during analysis.

Feedback activities in the local collective arena (Figure 6.2-2c) include comparing individual postfield results with related project datasets. Visualization tools may be used to discern data patterns and anomalies. Interaction between arenas and multi-dataset analyses represent positive feedbacks as datasets are formalized from preliminary versions to final states for release. For example, one project researcher may discover an error in a dataset when using another researcher's data. The use of data by other project researchers that take advantage of data made available in a local collective arena, serves as another check on measurements that contributes to the robustness of scientific data. Findings in the local collective area may be sent to inform project planning or may be published in papers and reports.

In addition to data analysis and data assembly in the local collective arena, other important activities take place. Documentation in the form of metadata creation was mentioned earlier. Transformational work with datasets may include identification, creation, and validation of new data products. For instance, several sources of weather data may be recast to the same time intervals to create a derived product of monthly means. Consistent calculation processes are developed. For example, a suite of images may be transformed into an index such as a calculation of extent of cloud cover in a region for local use or for new designated audiences (Baker et al., 2015a).

Project assembly and a public assembly points are shown on the left side of the local collective arena in Figure 6.2-2c. A local project data system with datasets and dataset artifacts supports these assembly points. In the project assembly arena with access available only to project members, versions of datasets are shared with the understanding that anomalies will be communicated informally between researchers. Delivery of data may be via a website or an online directory using file transfer protocol (ftp). A website often provides access to a data catalog, datasets, publications, and photographs. A project website is typically a collaborative undertaking for researchers that work together to describe the data sampling, the physical context of the data and the scientific context of the project.

The local collectives may establish relations with one or more external archives (Figure 6.2-3a) that provide data preservation and dissemination services. An archive's work is typically defined by an explicit mission as well as by formal procedures and practices such as a well-described submission information package (CCSDS, 2012). An archive may in turn have relations established with other archives (Figure 6.2-3b) so that metadata or data may reside in more than one archive or appear in more than one archive catalog. In Figures 6.1 and 6.2 provenance is clear when data moves from the project assembly to a designated archive and from there to alternative archives.

The rationale for portraying the movement of data within a Data Work System is to examine the dynamics across the six data work arenas including three portraying the local collective. Referring to local arenas simply as 'pre-archive, neglects the dynamics within the local arenas as sites of interactions about data for project researchers as well as between researchers and data specialists. The interdependence of work with datasets is evident as feedbacks shown as bi-directional arrows. For example, findings about data by one researcher may prompt reprocessing or reanalysis by another investigator of their datasets. Thus, the model does not compress pre-archive data work into a single or isolated activity. It is inclusive of both the data work of individual investigators in the local postfield (figure 6.2-2b) as well as of collective data work in the local component by researchers and data specialists that precedes submission of data to an archive (Figure 6-2c). Having described the Data Work System model, the next section discusses three findings from observing such a system: trade-offs, Data Gateways, and three kinds of Data Collectives.

6.3 Findings

The Data Work System model makes explicit the movement of data across data arenas depending upon both the data infrastructure within the local component and the relations developed with external archives. Each arena is a hub of activities that may include informal exchanges of ideas, information, and data as well as interactions with data systems. Arenas involve different degrees of socializing that contribute to collaborative work with interactions occurring at conceptual, theoretical, and practical levels. The system model, though still a simplification of data activities and interactions within a research project, serves to illustrate both the complexity of data work in terms of a variety of centers of activity and relations between the data work arenas. While vital to scientific data work, non-scientists may see the variety of feedbacks and ongoing dialogues as inefficient when they fail to take account of the dynamics and the constant cross-checking associated with data analysis. In many of the data work arenas, scientific inquiry and data work blend. Research routinely involves unanticipated data sampling, processing results, methods development, and/or documentation activities. Although some of the specialized processes are candidates for streamlining, local data work arenas are sites of innovative thinking where new approaches and unanticipated insights emerge.

Having studied the characteristics of data infrastructure earlier and modeled the data work system at the granularity of data work arenas that comprise each of the three system components in this chapter, a brief discussion of trade-offs is presented in the next section. The chapter ends with a discussion of Data Gateways and Data Collectives.

6.3.1 Trade-offs associated with elements of data infrastructure: A source of variation

A number of trade-offs were observed during this investigation associated with elements of data infrastructure in the local component (Sections 5.2.1-5.2.7). Table 6.1 gives some examples of trade-offs that make choices visible. Such trade-offs are discussed as transcontextual issues by Star and Ruhleder (1994) and as paradoxes by Lewis and Dehler (2000). The delineation of trade-offs increases our grasp of the kinds of choices that arise in data work and the decision-making that shapes the development of data work configurations.

One trade-off observed in all cases involves the kind of intermediaries supported to work with a project. For instance, the support for a systems administration group put AtmChem at the cutting edge of data exchange early on when data delivery in the 1970s depended on locally

supported shared storage devices using file transfer protocols and zipped files. In contrast, the LTER community norm of support for a dedicated data management position at EcoPrairie in the 1980s focused attention on local assembly of data, seeding growth of a Local Data Collective and development of a project data system.

Table 6.1: Trade-offs observed with elements of data infrastructure in the local component

Trade-Offs	Examples
Data and systems management	EcoPrairie full time information management team that participates in community-of-practice; AtmChem full time technical group that participates in community-of-practice
Knowledge and data production	EcoPrairie configured for knowledge and data production with embedded data manager; EcoRiver configured for knowledge production in traditional science model (Figure 5.1)
Local and external repositories	EcoPrairie using local data system repository; AtmChem using external archive (Figure 5.2)
Manual and digital processes	EcoPrairie gathering of data for local data assembly and automated submission via metadata validator to network sysem
Local and general representation	Locally designed project website and network standardized webpage for EcoPrairie
Dataset and project collections	EcoPrairie dataset migrated to network system and project collection sent to institutional repository
Short-term and long-term thinking	EcoPrairie local conventions use in project data workflow and capacity to generate standard metadata formats

Budget limitations in supporting data management and systems administration create what appears as an ‘either/or’ scenario or a trade-off, though the balance of data management and technical expertise depends also on available personnel and their skills. The interdependence of the two options of support for data work is shown in Figure 6.3a following an approach suggested by J. Cutcher-Gershenfeld (personal communication). Within a fixed budget, the focus of data work activities may be high in terms of support for systems administration and low in terms of support for data management as is the case for AtmChem. EcoPrairie had a full-time embedded data manager and less support provided for project-specific technical development.

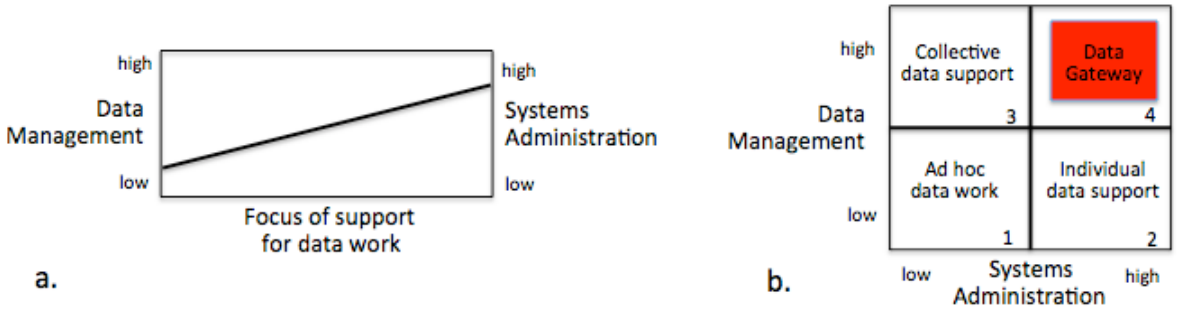

Figure 6.3: Trade-offs with intermediary roles chosen to support data work in the local component are shown in two views: a. the degree of support for data management in an inverse relation with systems administration; b. a two-by-two representation with four support options for data management and systems administration.

Choices about kinds of support may be viewed in an expanded four-quadrant landscape of data arrangements as shown by the alternate view of the relationship of data management and systems administration in Figure 6.3b. In this matrix, the first quadrant indicates ad hoc support for data work with little support for either data management or systems administration. With increasing systems administration as in quadrant 2, support for shared technical arrangements that enables individual investigators' work with digital files and applications is strong. With increasing project data management, support for collective data work grows (quadrant 3). When strong support is available in both data management and systems administration with the two interacting effectively (quadrant 4), conditions are established for development of a data gateway, an element of infrastructure discussed in the next section.

6.3.2 Data Gateway

Data gateway is a concept that arises in discussions of the development of technical systems and is also useful in describing an important aspect of the EcoPrairie project collective data work. In studies of cyberinfrastructure, gateways have been identified as creating path dependence (Edwards et al., 2007) and also as a phase of infrastructure growth (Edwards et al., 2009). While they have been studied largely in association with large scale technical systems as gateway technologies, scientific gateways, and gateway organizations (Wilkins-Diehr, 2007; Zimmerman and Finholt, 2007), in-situ design work has been discussed as part of 'work infrastructure' (Pipek and Wulf, 2009) and as 'in-between infrastructures' (Botero and Saad-Sulonen, 2010).

A Data Gateway for projects is defined here as an infrastructure element associated with the data work of a local collective that establishes and maintains relations with one or more archives. Interactions with external archives include data staging and/or submission as well as a constant consideration of local conventions and their relation to archive standards. A local Data Gateway is a sociotechnical bridge between data work that is internal and external to the collective. A collective with a gateway can be viewed as delaying or holding back data but from another perspective, as key to creating pools of data with data intermediaries supporting both local data management and local data infrastructure that enables discursive, deliberative, and integrative data activities.

6.3.3 Collectives

Three terms are used in the following sections to describe work with project data: Project Collective, Data Collective, and Local Data Collective. Project Collective is a general term that refers to project members working individually and collectively on data related to the project. For the assembly and collective management of project data in a particular location, Data Collective is used to designate a project's members working together on project data with support of data specialists in local and/or non-local data arenas. A Local Data Collective is one type of the Data Collective that occurs when project members and specialists work to coordinate and manage project data and a project data system exclusively within the local component.

Table 6.2 summarizes the kinds of Data Collectives for the three cases. The first two rows in the table contextualize the projects in this study and are followed by four characteristics related to services: data assembly tools, data assembly focus, digital data objects, and data workers. Two salient characteristics highlighted in the cross-case analysis are added to further delineate the project data collective: internal relations described by the science-data work models (section 5.2.5, Figure 5.1) and external relations described by the configurations of the cases (section 5.2.6, Figure 5.2).

The first project Data Collective is called a Local Data Collective. A key element of the collective is a data system that assembles and manages project data using a mix of loosely and highly structured processes. The collective effort is sensitive and responsive to changes in data and technology as well as social circumstances of local data arenas that support local data use and data production. Data work in this collective project effort involves continuing design,

adaptations, and modifications that address 'break-downs' and spur (re)design of functionality internally within a local arena and externally in relations with data archives. As a mesoscale way station, the Local Data Collective is where data resides prior to its archive in external data facilities. There is some recognition of the benefits to considering these design testbeds for data assembly and data system development (Jensen and Morita, 2016). Data assembly is coordinated by and situated at a project data office that supports a project data server. Data workers support scientific data work, internal relations of the collective, and external relations as a gateway for data to be moved to archives. Data work involves a variety of digital data objects that are project vetted and assembled on the website in what constitutes a pre-archive project collection. This collective is pro-active in identifying community-specific data practices, responding to data needs, and designing new supports as well as maintaining existing project data infrastructure by adjusting, augmenting and redesigning elements and conventions. It provides an example of data infrastructure arrangements that contributes to development of science data practices, local data conventions and data archive processes. The Science with Embedded Data Management model or the Science with Partnership model describes internal data relations for a Local Data Collective. External relations with the archives are handled via a project Data Gateway.

Table 6.2: Kinds of Data Collectives for the three cases

#	Data Collective Characteristic	Local Data Collective (EcoPrairie)	Developing Data Collective (EcoRiver)	Archive Data Collective (AtmChem)	Data Collective Category
1	Sampling context	single location	single location	multiple locations	context
2	Existing infrastructure	local technical support single project data infrastructure data management community-of-practice organizational IT group institutional repository	local technical support organizational IT group	local technical support technical community-of- practice organizational IT groups organizational archives	context
3	Data work focus	local collective	conceptual development	project collective	services
4	Data assembly approach	data systems project website	in planning	archive data system, project landing page	services
5	Digital data objects	data files, datasets, data packages, pre-archive collection	in planning	dataset collection, project collection	services
6	Data workers	researchers, local data specialists, library data specialists	researchers	researchers, archive data specialists, library data specialists	services
7	Internal work relations	Science With Embedded Data Management	Traditional Science	Science with Data Production	relations
8	External relations with archives	varies from none to data system submit	in planning	individual data submit	relations

EcoPrairie is an example of a Local Data Collective with a Data Gateway (Figure 5.2c). The EcoPrairie data intermediaries while fostering both collective and integrative data activities in local data arenas, were mindful of the importance of continuously making decisions about how best to balance at any one-time science needs, archive requirements and options, as well as changes in technology. With a position established to manage local data assembly, issues such as data capture, formatting, protocols, quality, and documentation were addressed as primary tasks. Metadata creation, dataset formatting, and data delivery were also managed through time consuming processes involving elicitation and judgments about data carried out in conjunction with project researchers.

EcoRiver is an example of a Developing Data Collective that responded to a funding agency requirement for data access in 2013 and subsequently held a data stewardship workshop to further collective data work. Initial data management activities included writing a data management plan for an agency proposal, a formal data management job position description, and a vision statement. It describes a project in the process of identifying digital approaches for assembling data collectively by developing conceptual tools and vocabulary for local data work as well as pilot data efforts. Members may have direct relations with other investigators or archives. Data assembly may be piloted using publicly or organizationally available storage media and data objects are largely loosely structured data files. The Traditional Science model describes this collective's internal data relations. When external relations with archives are established, they are handled by direct interaction between individual investigators and a project-designated archive.

AtmChem is an example of an Archive Data Collective where members submitted their datasets directly to a remote archive. This third kind of Data Collective carried out joint sampling in many locations. AtmChem worked with science field project leads for each set of field campaigns. In addition, it coordinated long-term within its organizational unit that supported informal data storage and sharing. For this case, integrative data work that began in the field extended to modeling efforts. Data was assembled in collaboration with an archive that maintained data systems undergoing continuous design. Data assembly often began in the field with an archive supporting fieldwork. For this configuration, archive staff filled the role of data manager, designating a data contact for each field project as liaison between the archive and the research effort. The Science with Data Production model describes internal data relations for this

collective. External relations with archives were handled by direct interaction between individual investigators and a project-designated archive.

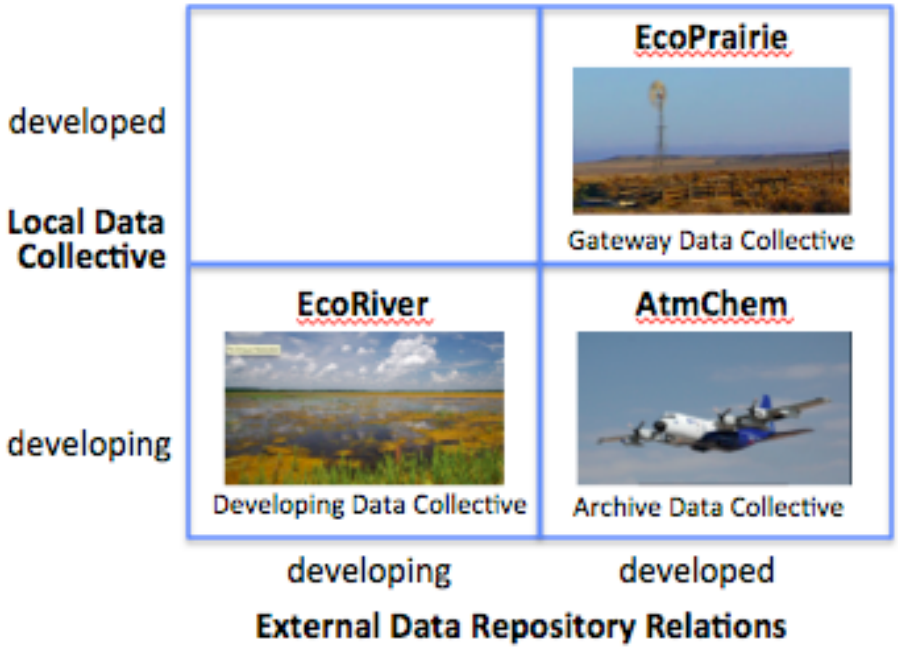

Figure 6.4: Four quadrant presentation of the three cases that portrays

A snapshot for the three cases in their current configurations is provided in Figure 6.4 where two of the elements of data infrastructure are highlighted. EcoRiver is shown as 'developing' both in relations with external archives and with its collective data work (Figure 5.2a). AtmChem is shown as 'developed' in its relations with archives. Since researchers had individual responsibility for data production and for submitting data directly to the archive, the local collective has less of a role in assembling the data and is shown as 'developing'. EcoPrairie, a Local Data Collective with a Local Data Gateway and a team of information managers working with researchers on data production, has established relations with external data repositories (Figure 5.2c).

In summary, a model was developed with three arenas detailing activities in the local component that together with a field arena and two archive arenas constitute a research data work system. This representation of the data work configuration allowed the complexity of data work in the local component to be investigated in relation to archiving. Findings included some

of the trade-offs that contribute to development of configurations and the concept of a Data Gateway associated with a Local Data Collective. Three types of Data Collectives were identified: Developing Data Collective, Local Data Collective, and an Archive Data Collective. In the next chapter, a review of findings and the research questions is presented followed by brief discussions of changing data cultures and future research.

CHAPTER 7. CONCLUSION

With the advent of the Digital Age and the prioritization of data access, it is often difficult for scientists and non-scientists alike to comprehend the many factors involved in collective data work and the production of data to be archived. A Data Work System model illustrates why descriptions of data work and Data Collectives are elusive. Though often absent from large-scale information infrastructure architecture plans, mesoscale data infrastructures play a significant role today in supporting both scientific collaboration and data production.

By investigating the characteristics of data work configurations in ongoing scientific research projects, this study highlights the role of mesoscale infrastructures as a source of support and flexibility for data work critical to scientific inquiry. A systems lens captured the dynamics of data work drawing attention to the human agency involved in data work including support for feedbacks and cross-checking of data, activities subject to rapid revision, close scrutiny, and interpretive finesse.

Case studies provide a small but important window into data work in the field-based sciences. Their stories deepen our understanding of scientific collaboration and information environments. Case studies motivate, inspire, and contribute through illustration (Siggelkow, 2007). The cases studied illustrate the significance and variety of data work configurations. Motivation for the cases in this study is twofold: to open up discussion of configurations that support scientific projects and to increase awareness of mesoscale data infrastructure. By highlighting the complex support arrangements as elements of a dynamic data work system, the cases are intended to inspire description of additional configurations that inform the development of data infrastructures.

A summary of findings is provided below followed by a section on the changing nature of data culture. The last two sections present future research and some final thoughts.

7.1 Summary of Findings

Drawing on the findings in previous chapters, responses to the research questions are summarized in this section. The summary ends with a cautionary tale about the need for sensitivity to the range of infrastructures that include, among others, large-scale and mesoscale.

Research Question 1

In response to the first research question about how differing configurations of scientific research projects and data repositories support the movement of data, an analytic framework was used to capture the full context of data work configurations, thereby accounting for field, local, and archive components (Table 5.1). In addition, development of a Data Work System model inclusive of a variety of data work arenas allowed the number and arrangement of arenas to be customized to the specificity needed within the system (Figures 6.1, 6.2). The use of any number of data work arenas allows the model to be used to zoom in on any designated part of the system. For instance, a research project focusing on the field component could identify a number of data work arenas associated with fieldwork. For this study only a single, token field arena was identified since the focus was on the local component. The ability to change the granularity of a study's focus is an important feature that lends itself to investigation of alternative configurations whether in the ecological and atmospheric sciences or in other field-based sciences.

A related research question asks how project-repository configurations differ. Project Collective relations with external archives are illustrated in Figure 5.2. Use of the framework and the model, enabled investigation of the local component and existing mesoscale infrastructures for each case. Three kinds of Data Collectives were identified that support the movement of data from projects to data archives: 1) Local, 2) Archive, and 3) Developing Data Collectives.

EcoPrairie, a Local Data Collective characterized as a way station situated within the local component between the field and archive components, supported the project's local information environment as a dynamic space conducive to the generation of knowledge. The collective functioned as a Data Gateway with its data system, data manager, and local data conventions. As an assembly point for data, the project illustrated the characteristics of a Local Data Collective, a particular kind of Data Collective with data expertise available within the local component. The Data Gateway was closely aligned with two external data archives.

AtmChem, an Archive Data Collective, supported assembly of project data at a facility external to local arenas. It featured close coordination of project researchers and archive staff on prefield data planning, field support, and data submission. Typically a member of the archive was designated as a data management contact for each project. When project researchers submitted multiple versions of the data, early versions were made available only to project

members. Development of a project website by archive staff drew together into a standardized template all the aspects of the project work including project datasets, project artifacts, and other project-related efforts.

EcoRiver, a Developing Data Collective, involved project researchers sharing data informally without project collective relationships with archives. Researchers were aware of the need to develop procedures for collective data work; they were in the process of developing a conceptual understanding of and a vocabulary for discussions of collective data management and open access. These concepts were recognized as requiring incorporation into their professional and institutional visions as well as into their personal and collective data practices. A variety of circumstances influenced this project's approach to change including the initial activities that resonated for the collective, the kinds of partnerships involved, in addition to the institutional support and infrastructures available. Given the potential to change in many arenas, it was difficult to foresee next steps in the configuring of their collective data efforts.

In considering project and repository relations, a number of similarities in the data work configurations were observed. The cases involved partners from a variety of sectors, often working at different scales including fieldwork, project, institutional, and network. In addition, many were actively learning about concepts and vocabulary new to them (e.g. data archive, data stewardship, data curation) and about new conventions for data assembly (e.g. datasets, data packages, project collections). Further, all the projects were in the midst of a major event. Termination of EcoPrairie illustrated the role performed by the project data manager at closing in archiving project datasets and in tailoring closing activities to both immediate and future needs. At EcoRiver, the recent agency requirement for data management plans coupled with the award of two grants led to introduction of the concept of data management and a pilot project on data assembly. AtmChem, while planning to provide expanded community support by developing a new center, continued field campaigns and developed a local catalog.

Differences that distinguish and influence configurations were also observed. One major difference involved the difference in fieldwork logistics. Each collective performed collaborative research and conducted field sampling using a shared platform at a designated location. The two ecological cases (EcoPrairie and EcoRiver), however, sampled and shared information from the same geographic site in repeated visits to the field with differing frequencies, e.g. ad hoc, daily, seasonally, and annually. The platform-based AtmChem carried out many field campaigns in

different locations using aircraft as a shared platform that required advanced technical support and involved instruments sampling at different rates. Each campaign can be considered a subproject in the overarching AtmChem mission to support community efforts that target significant sampling locations. AtmChem worked with several data facilities that provided both field support and archive support.

For EcoPrairie participants, a Local Data Collective created a familiar environment that provided researchers and data workers alike everyday experience with collective data work and data issues. The processes of data assembly and data-related documentation prompted learning and represented knowledge-making opportunities for project members. Collectives served as an effective training ground for the many kinds of data workers. The Local Data Collective provided exposure to and sometimes integration of different perspectives on data and technical constraints and trade-offs integral to data work. This prepared project participants to better understand the language and issues associated with data work and information architecture as well as the ramifications of data-related decisions. Ongoing interactions within collectives enabled intermediaries to design and tend the local information environment, tailoring its development to project needs. Tasks involving data catalogs, access, review, and visualization became part of a project's information environment. In addition, by designing and tending data workflows, intermediaries provided the human support for individual project members awash in new technology and digital options.

The availability and maturity of existing infrastructures impacts the development of local data work arenas. For instance, the EcoPrairie Local Data Collective had a number of local infrastructure supports available from project and GIS teams, an organizational unit, a campus-wide IT unit, an institutional repository, and a data management community of practice. The AtmChem Archive Data Collective had significantly different infrastructures available: local technical support, several large organization-wide IT groups including those with archives, and a technical community of practice. The EcoRiver Developing Data Collective had three mesoscale infrastructures available: local technical support, a partner's web master, and a small organization-wide IT group for the state. The differences such as EcoPrairie interacting with a data management community of practice and AtmChem with a technical community of practice are captured in Table 6.2.

Research Question 2

In response to the second research question about the elements of project data infrastructure that contribute to the movement of data to repositories for field-based research projects, this analysis focused on support in local data work arenas. Eight elements describing data infrastructure for collective data work were identified: data management, systems management, data assembly including data systems, website including an information environment, internal relations of science and data work, external relations with archives, data collections, and data intermediary roles. These elements, described in detail in Chapter 5, support data work within arenas as well as between data work arenas.

In considering how Data Collectives vary, characteristics were identified and summarized in Table 6.2 for the three cases. The first two characteristics describe the project context and existing infrastructure as mentioned above. Four service characteristics identified include data work focus, data assembly approach, digital data objects, and data workers. The data work focus highlights whether assembly of project data is a priority, and the data assembly approach captures whether a data system developed in the local component. Digital data objects are typically grouped for archiving in one of two ways: in a dataset collection consisting of data and metadata or in a project collection consisting of datasets and other project-related artifacts. The data worker category includes researchers and data specialists. The last two characteristics have to do with relations: internal data work relations (Figure 5.1) and external relations with archives (Figure 5.2) that have been discussed and presented together visually in Figure 6.4.

Considering the elements of project data infrastructure and their translation into characteristics of Data Collectives, makes evident the role of data intermediaries as data workers in a variety of data work arenas. Consequently, this investigation addresses not only the what, where, and how of data work but also the 'who' of data work. Data intermediaries associated with collectives provided ongoing support that ensured the design, functionality, and growth of mesoscale data infrastructures in particular. The role involved ongoing mediation within the research environment as well as in meeting the changing circumstances within the digital environment. Key intermediaries contributed to the conceptual development of data work, taking into account changing technology and data practices. They contribute to the resiliency of data work not only by their continuing attention to local data needs but also by their infrastructural thinking.

Finally, in considering why data infrastructure and attendant data work differs, initial conditions and Project Collective history such as funding arrangements and project events made a difference. Both EcoPrairie and AtmChem evolved from quite different initial conditions. Eco-Prairie was launched with attention to data work ensured by a local data management position that focused on data assembly as well as the growth of local infrastructure. EcoPrairie as a Local Data Collective began data assembly at a time when little data infrastructure existed and there were few domain archives. As a result this Local Data Collective developed the technology and data practices needed to assemble project data locally. AtmChem, an Archive Data Collective residing in a national data center, was initially planned to support collective sampling and to partner in scientific collaborations with the atmospheric science community. External field support, needed for its shared platform and changing research teams, included growing attention to collective data and archive work.

Funding for EcoPrairie and AtmChem was a combination of both core funding and grant-awarded field-project supplementary funding. The EcoPrairie project and science offices as well as its department changed with the project's six-year cycles. While the science offices of field-project campaigns of AtmChem changed, its archive partners had stable locations within a long-term organization though they faced the disadvantage of having to bridge the gap between science-based and technology-based efforts.

Once initiated, collective data work varied depending upon the priority given to various infrastructure elements. The Gateway and Archive Data Collectives differed in terms of trade-offs, that is, the choices made with respect to the elements of data infrastructure. Project researchers grappled in practice with trade-offs involving geographic and temporal extent as well as project scope and design complexity (Table 6.1). Collectives managed the equivocality and uncertainty that enter in fieldwork with the selection of a finite number of measurements to represent the natural world, of options for information architecture, of constraints that exist in using digital data systems to organize data, and of options for developing data products.

A cautionary tale

Awareness of the role of data collectives in supporting scientific inquiry will shape the role they play in the scientific data landscape. In addition to being a narrative about the variety of data work configurations and the existence of mesoscale data infrastructure, this research is a

cautionary tale about scientific data work and the 800-pound gorillas:

> My impression was, at least the attendees list was dominated by the large data generators. Satellite people and global modelers … the enormous data generators or EDGs, they know what they are doing with terabytes. They know how to store data in one hemisphere and get into their laptop in seconds in a different hemisphere. The people who really need the help are the day-to-day experimentalists who produce one number at the bottom of a spreadsheet that somehow may be more valuable than terabytes of model output. … So each measurement is unique in its own. And I guess that's what they call the 'tail end'. I have some real concerns about the 800-pound gorillas not giving a chance to the tail end. (UM)

The 'tail end' in this context refers to data collections by size with large homogeneous collections in contrast to an assembly of many smaller collections that often involve heterogeneous kinds of data (Palmer et al., 2007; Heidorn, 2008). The case studies describe mesoscale infrastructure that supports 'tail end' heterogeneous data generators as distinct from infrastructures for large-scale, homogeneous data streams of 'enormous data generators'. Long-tail case studies are beginning to reveal characteristics of long-tail scientific data activities (e.g. Darch et al, 2015; Ferguson et al, 2014; RIN, 2009; Karasti et al, 2006; Karasti and Syrjänen, 2004). Making distinctions in kinds of data work configurations and infrastructure is a necessary step in designing for diversity. Designing for multiple scales and scopes addresses the divide that typically separates archives from local data work arenas, making room for claims of efficiency for centralized systems in theory *and* for accounts about of effectiveness of Local Data Collectives in practice.

Given the importance of post-field and pre-archive data work, scientific inquiry requires plans for highly structured 'hubs and spokes' as well as for a less structured web of repositories supported by mesoscale infrastructures (Baker and Yarmey, 2009). Harvey et al (2017) address infrastructure themes such as development, environmental, and digital, raising questions about connectivity and gaps, governance and control, as well as the possibilities of alternative configurations. They speak to the potential for development of digital infrastructures that centralize and administrative structures that integrate to erode local autonomy. As an initial assembly point for project data, Local Data Collectives counterbalance the large-scale through development of mesoscale infrastructures that enable a web of repositories not only as points for data assembly and for growth of information environments but also as venues for change and

prototyping.

7.2 Changing Data Culture

In each of the cases studied, collective data work was in a state of change - changing technology, scientific collaborations, organizational arrangements, and cross-sector partnerships. The cases involved local and network scales (e.g. project website and network-level standardized content), current and legacy science (e.g. primary science web page and legacy landing page), tightly and loosely structured data work (e.g. systems support and data management support), and archive of collections (dataset and/or project collection).

Amidst the change, a research scientist explained: "You know, I have come around to the idea that if it's not available publicly online, it doesn't exist" and pointed out that this was a change from past practices prompted by the complexity of questions now being addressed:

> In the mid 90s, I started out when you collected your data and you saved your data and you worked with your data. But I've seen it then transform that to smaller groups . . . so I think science in general has changed, you used to have to build your own kingdom. … Now the scientific community is much more understanding that the only way for us to understand these complex problems is to work together and share data. … As I got more involved in the LTER I saw the benefits of network science. Not necessarily to myself but to the community … I think, I saw a better way of doing science. You could share resources. You could share intellectual capital. There were a lot of things that were shared that would not be shared in the 'my kingdom' sort of laboratory – the my lab, your lab kind of thing. (RS)

In speaking about the changing data culture, one participant emphasized the lack of a single solution:

> It's not a one size fits all by any means. And that's what I'm learning as I learn more and more about research data management. That's one of the issues that makes it so complex. There's not a one size fits all by any means. [laughter] And a lot of times it depends on the discipline and type of data and size. (UM)

Within programs developing mesoscale infrastructure, change is recognized explicitly in terms of the changing nature of scientists (Willig and Walker, 2016) and of the culture (Mauz et al., 2012). One research scientist reported: "I think working with the data issues at LTER has really fundamentally changed my perspective … about collaboration and about the role of data

management." Further, in terms of organizations, the aim to coordinate rather than to control is noted:

> We have a data management committee that was charged and it's a standing committee that came out of three data management forums held over the last three years. … But, you know, it coordinates the data environment, it doesn't you know control it. And so that's consistent with our distributed culture. (RM)

Interviews with research participants revealed that the digital age has not only brought discussions of change but also responses involving new perspectives on collaboration and collective data work. Having underscored the changing data culture, future research is addressed in the next section.

7.3 Future Research

Once understanding within data cultures includes recognition of the complexity of data issues and to express cautionary tales about the need for sensitivity to diversity, then possibilities open up for re-envisioning data work configurations as varied and responsive to scientific needs. Characterizing and developing mesoscale infrastructures contributes to this effort. There are a number of directions to pursue with future research including additional case studies, development of transparency through documentation, the differing temporal scales relating to data work, and the roles of intermediaries in various data work configurations.

Additional ethnographic case studies are needed to gauge the existing variety of mesoscale data work configurations. More case studies will provide additional information about Data Collectives. Investigations in field sciences other than the ecological and the atmospheric sciences will reveal how well the eight elements of data infrastructure for collective data work describe their configurations. They would provide some indication of the generalizability of the Data Work System model. Additional case studies are needed to provide further information about trade-offs. Finally, longitudinal ethnographies will permit investigation of patterns of evolution of data collectives.

Qualitative approaches can continue to contribute to investigation of data work configurations. The elements of data infrastructure for collective data work in Table 5.3 and the characteristics of Data Collectives in Table 6.1 can be used in developing future interview

instruments that elicit further information about factors influencing choices made in local data work arenas. Case studies with ethnographers as coparticipants will contribute to preparing project researchers for decision-making relating to the trade-offs encountered in planning the growth of data infrastructure and with development of project-related information environments. Such studies are particularly beneficial when considered in light of the mutual learning reported in the stories of participation that fostered awareness, stirred reflection, and supported conceptual development (Appendices B.2.6, C.2.4, D.3.4). Coparticipation increases the number of individuals – researchers, data specialists, and intermediaries – prepared for collective data work and the design of data infrastructures.

Transparency is often an overlooked issue in information environments. Transparency, referring here to the public availability of information about data work, is key to design that facilitates learning. Lack of transparency often occurs when a system's ease of use takes priority and precludes ready availability of all the details relating to the organization of data and data work. For instance, one case revealed a tradition of providing services that make it easier for project researchers to retrieve resources without exposure to a clutter of information about packaging and delivery. As a result, metadata was used for display purposes but not made available in its entirety to those interested in fully understanding the organizational scheme in order to make use of it or to contribute to data activities. For example, full disclosure of metadata available about a dataset contrasts with a singular focus on internal use but not display of the metadata. Full disclosure means making available information about the processes of data generation and delivery via online postings including keyword lists and data workflows. We know little about how transparency is maintained in practice given time, personnel, and budget constraints or about the role of transparency in the development and functionality of adaptive systems.

Time and timing loom as significant issues in design, and are frequently focused on arrangements within existing structures. Time is largely absent from the eight elements of local data infrastructure summarized in Table 5.3. Incorporation of multiple perspectives and rhythms relating to time is needed in planning data infrastructure (e.g. Karasti et al., 2010; Jackson et al., 2010). The co-construction of timelines for the cases provided a sense of the context and change over time in circumstances within which collective data work occurred. The mix of major and

minor events impacting data work became evident as timelines were expanded and subset depending on which issues prevailed at the moment.

Finally, an expanded investigation of the roles of intermediaries in collective data work can be informed by the concepts of community, participant, and expertise. Research on Project Collectives informed by this trio of concepts would provide insights into researching and working with project researchers on complex data work systems and configuration. These concepts are broad and multi-faceted. Previous research has established differing views and definitions of them. Highly relevant to data work are the multiple forms of learning in communities such as Communities of Practice (Lave and Wenger, 1991), the multiple forms of participation including staged participation and design-in-use (e.g. Saad-Sulonen, 2013), the multiple forms of intervention such as midstream modulations (e.g. Fisher et al., 2006), in addition to the multiple kinds of expertise such as interactional expertise featured in a 'periodic table of expertises' (e.g. Collins and Evans, 2008). A table of brokerage terms - liaisons, mediators, boundary spanners and/or brokers of knowledge and information – provides an overview of kinds of participation and motivations (Long et al., 2013). Investigating the ensemble of community, participation, and expertise concepts provides a point of departure for considering the distribution of data work responsibilities. New kinds of data work require new kinds of data workers in intermediary roles.

Having considered some potential future research directions of research, a few concluding thoughts are given in the next and final section.

7.4 Concluding Thoughts

In seeking out and learning from existing project configurations and data work arenas, there is an opportunity to understand how mesoscale data infrastructures and Data Collectives has developed in practice to handle new kinds of data and data needs. There is a tendency to take an overly optimistic view of information architecture in spite of the complex design issues involved. Venturi (1966; see epigraph) created a 'Gentle Manifesto' as a reminder about work that blends the logic of construction with the art of design. A Gentle Manifesto for data work would speak to situated diversity, local trade-offs, and the uncertainties inherent to scientific

inquiries. Data workers and their contributions to Data Collectives would be recognized as critical in discussions of digital architectures as roadmaps mature and are operationalized.

The findings of this study illuminate Data Collectives within existing data work configurations. The identification of three Data Collectives has shown us the need to inquire further about additional kinds of data work configurations and the growth of local data infrastructure. Data workers associated with Data Collectives ensure responsiveness to local data needs, support for the knowledge work of science, management of data production, and coordination with archives.

Mesoscale infrastructures, Project Collectives, and Data Gateways bridge the gap between generation and preservation of scientific data. In considering data work, as stated by Star (2000), “it's infrastructure all the way down”. But of course, with data from the field moving to archives, “it’s infrastructure all the way up”. And lastly, from the perspective of those at mesoscale way stations, it’s interdependent infrastructure all the way through.

APPENDIX A. Institutional Review Board Instruments

A.1 Acceptance Letter

UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

Office of the Vice Chancellor for Research

Institutional Review Board
528 East Green Street
Suite 203
Champaign, IL 61820

April 5, 2013

Carole Palmer
GSLIS
501 E. Daniel Street
Champaign, IL 61820
MC-493

RE: *Data Sites: Collaborative Practices and Data Repositories*
IRB Protocol Number: 13595

Dear Dr. Palmer:

Your response to stipulations for the project entitled *Data Sites: Collaborative Practices and Data Repositories* has satisfactorily addressed the concerns of the UIUC Institutional Review Board (IRB) and you are now free to proceed with the human subjects protocol. The UIUC IRB approved, by expedited review, the protocol as described in your IRB-1 application with stipulated changes. The expiration date for this protocol, UIUC number 13595, is 04/04/2014. The risk designation applied to your project is *no more than minimal risk*. Certification of approval is available upon request.

Copies of the attached date-stamped consent form(s) must be used in obtaining informed consent. If there is a need to revise or alter the consent form(s), please submit the revised form(s) for IRB review, approval, and date-stamping prior to use.

Under applicable regulations, no changes to procedures involving human subjects may be made without prior IRB review and approval. The regulations also require that you promptly notify the IRB of any problems involving human subjects, including unanticipated side effects, adverse reactions, and any injuries or complications that arise during the project.

If you have any questions about the IRB process, or if you need assistance at any time, please feel free to contact me or the IRB Office, or visit our Web site at http://www.irb.illinois.edu.

Sincerely,

FOR

Anita Balgopal, Director, Institutional Review Board

Attachment(s)

c: Karen Baker

Informed Consent for Observation
UIUC Data Sites

Thank you for your interest in participating in this study. Our main objective in this research is to investigate data practices, data sharing, and interfaces with data repositories in the natural sciences This research is led by Carole L. Palmer, Director of the Center for Informatics Research in Science and Scholarship at the University of Illinois at Urbana-Champaign in the Graduate School of Library and Information Science. Karen S. Baker is a graduate student working with Dr. Palmer. If you agree to participate in this research, you are contributing to the general knowledge of organizing, managing, and packaging data.

If you agree to participate, observations will occur during meetings and group activities as well as arranged following an interview or in the course of your research work. Notes will be taken on discussion and communication of data and scientific practices and issues. Observation sessions will be scheduled. They will occur generally from one to twelve times over a period of one month to a year. Duration of an observations will vary from approximately a half hour to all-day for community meeting or will vary from approximately a half hour to two hours for informal group meetings or meetings with individuals.

We may ask to record these sessions using a digital audio recorder. The sessions will be recorded only if everyone involved in the session agrees to the recording. We may also ask you to share copies of materials relating to data management, and for your permission to take photographs of relevant workspace and data materials. Consent will be obtained from all those participating, and no compensation will be made to individuals participating in this study. All contributions will remain confidential unless you provide express written information otherwise. Participation is entirely voluntary and you may stop at any time. You can skip any question you don't wish to answer with no negative consequences. There are no risks involved in participating in this research beyond those risks that exist in daily life.

You will be given a copy of the consent form for your records. Your decision whether or not to participate will not affect your future relations with the University of Illinois at Urbana-Champaign. You are under no obligation to participate in the study. You are free to (a) end your participation in an observation session at any time, (b) request that the audio recorder be turned off at any time, (c) skip any questions that you do not wish to answer, and (d) request that your contributions to a recording session be destroyed and excluded from the study. If you have any questions, please contact Karen S. Baker at (858) 361-1158, ksbaker2@illinois.edu or Dr. Carole Palmer, at (217) 244-0653, clpalmer@illinois.edu. If you have any questions about your rights as a participant in this study or any concerns or complaints, please contact the University of Illinois Institutional Review Board Office at 217-333-2670 (collect calls accepted if you identify yourself as a research participant) or via e-mail at irb@illinois.edu.

By signing below you verify that you have read and understood the information provided above, that you are 18 years of age or older, and voluntarily agree to participate in this study.

_____

Signature of Participant

Please answer the following questions by checking off the yes/no responses and by signing your initials:
I agree to audio recording during this observation.

☐ Yes _____Initials ☐ No _____Initials

I grant the investigator permission to use excerpts of the transcripts from the audio-recording during observations in peer-reviewed journals, reports, and academic conferences as well as a dissertation.

☐ Yes _____Initials ☐ No _____Initials

I grant the investigator permission to use photographs of my equipment and work areas as well as scientific activities and site landscape in peer reviewed journals, reports, and academic conferences as well as a dissertation.

☐ Yes _____Initials ☐ No _____Initials

I grant the investigator permission to use photographs of me in the context of this study in peer reviewed journals, reports, and academic conferences as well as a dissertation.

☐ Yes _____Initials ☐ No _____Initials

I agree to allow my identity to be linked to my responses in peer reviewed journals, reports, and academic conferences as well as a dissertation.

☐ Yes _____Initials ☐ No _____Initials

_____

Date

Informed Consent for Audio-Recorded Interview
UIUC Data Sites

Thank you for your interest in participating in this study led by Carole L. Palmer, Director of the Center for Informatics Research in Science and Scholarship at the University of Illinois at Urbana-Champaign in the Graduate School of Library and Information Science. Karen S. Baker is a graduate student working with Dr. Palmer. Our main objective in this research is to investigate data practices, data sharing, and interfaces with data repositories in the natural sciences. If you agree to participate in this research, you are contributing to the general knowledge of organizing, managing, and packaging data.

Participation will involve a telephone or face-to-face interview. The interview(s) will take approximately 60-90 minutes, and held in your office or in a space of your choosing nearby. Follow-up interviews (typically 1 or 2) of approximately 60-90 minutes will be requested if clarification is required. During these interviews you will be asked approximately 10-20 questions about your research area, data activities such as collecting, managing, and sharing data, and interactions with data repositories. All answers will remain confidential unless you provide express written permission otherwise. Participation is entirely voluntary and you may stop at any time. You can skip any question you don't wish to answer with no negative consequences. There are no risks involved in participating in this research beyond those risks that exist in daily life.

You will be given a copy of the consent form for your records.. Your decision whether or not to participate will not affect your future relations with the University of Illinois at Urbana-Champaign. If you have any questions, please contact Karen S. Baker at (858) 361-1158, ksbaker2@illinois.edu or Dr. Carole Palmer, at (217) 244-0653, clpalmer@illinois.edu. If you have any questions about your rights as a participant in this study or any concerns or complaints, please contact the University of Illinois Institutional Review Board Office at 217-333-2670 (collect calls accepted if you identify yourself as a research participant) or via e-mail at irb@illinois.edu.

By signing below or agreeing verbally on the recording, you verify that you have read and understood the information provided above, that you are 18 years of age or older, and voluntarily agree to participate in this study. You are also agreeing to be audio-recorded during your in-person or phone interviews. However, at any time during the interviews, you retain the option to end your participation, at which time any recordings of you will be erased. If you do not wish for the interview to be recorded, please indicate so, and we will only take notes by hand. You are under no obligation to participate in the study. You are free to (a) end your participation in the study at any time, (b) request that the audio recorder be turned off at any time, (c) skip any questions that you do not wish to answer, and (d) request that a recorded session be destroyed and excluded from the study

_____

Signature of Participant

Please answer the following questions by checking off the yes/no responses and by signing your initials:

I agree to the audio recording of this interview.

☐ Yes _____Initials ☐ No _____Initials

I grant the investigator permission to use excerpts of the transcripts from the audio-recorded interview with organizational and//or institutional information retained in peer reviewed journals, reports, and academic conferences as well as in a dissertation.

☐ Yes _____Initials ☐ No _____Initials

I grant the investigator permission to use photographs of me and my work areas in peer reviewed journals, reports, and academic conferences as well as in a dissertation.

☐ Yes _____Initials ☐ No _____Initials

I agree to allow my identity to be linked to my responses in peer reviewed journals, reports, and academic conferences as well as a dissertation.

☐ Yes _____Initials ☐ No _____Initials

_____

Date

University of Illinois
at Urbana–Champaign

Institutional Review Board Office
528 East Green Street, Suite 203, MC-419
Champaign, IL 61820
tel: 217-333-2670 fax: 217-333-0405
E-mail: irb@illinois.edu Web: www.irb.illinois.edu

WAIVER OF DOCUMENTATION OF INFORMED CONSENT (45CFR46.117(C))

ALL APPLICATIONS MUST BE TYPEWRITTEN, SIGNED, AND SUBMITTED AS SINGLE-SIDED HARD COPY. PLEASE, NO STAPLES!

Responsible Project Investigator (RPI):

Last Name: Palmer	First Name: Carole	Dept. or Unit: Grad School of LIS
Phone: (217)244-0653	Fax: (217)244-3302	E-mail: clpalmer@illinois.edu

Project Title:

Data Sites: Collaborative Practices and Data Repositories

To request a waiver of documentation (signature) of informed consent, please provide a response to EITHER of the following questions. Please be specific in explaining why either statement is true for this research.

(1) That the only record linking the subject and the research would be the consent document and the principal risk would be potential harm resulting from a breach of confidentiality. Each subject will be asked whether the subject wants documentation linking the subject with the research, and the subject's wishes will govern. *Note: A waiver of documentation of informed consent is **not permissible under this category if the research is subject to FDA regulation.**

(2) The research presents no more than minimal risk of harm to subjects and involves no procedures for which written consent is normally required outside of the research context. **

This is a minimal risk study. There are no known physical or psychological risks associated with participation in this type of study. None of the activities associated with the study are designed to cause stress or fatigue. The work is of a discriptive nature about existing practices and emergent processes and policies. Multiple views are expected since organizations do not in general have data production policies in place. However, scientists may still perceive a risk. They may feel their opinion differs from that of others in positions of authority or that information gathered in our study could reflect on their ability to meet the stated goals of the project. We will engage procedures to minimize these risks such as keeping the data and identities confidential when requested in dissemination materials. Note: if the school were asking for this information to improve services, informed consent would not be required.

Requested only for phone consent

** In cases in which the documentation requirement is waived, the IRB may require the investigator to provide subjects with a written statement regarding the research.

RPI Signature: Carole L. Palmer Date: 4/5/2013

IRB Member Approval: _____ Date: _____

APPROVED
APR - 5 2013
UIUC INST REVIEW BOARD

062010 rev

A.2 Interview Instruments

Interview Guide – Data View

- Study explanation: This study is about data - data practices, data sharing, and interfaces with data repositories in the natural sciences.
- Request signing of permission forms.
- Turn on recorder.

I. Data Practices – Data practices refer to the customary way you handle and work with data.

1. Please describe the kinds of data with which you work. Is this data that you collect yourself?

2. What are the different stages or transformations that your data go through from the time of collection until the time it is ready to be shared? Describe one example that might be considered typical and one example that might be considered unusual or specialized.

3. In your last move to a new institution/position, which data did you bring with you? How did you decide? If you were moving to new facilities/department, which data would you take with you and why?

4. Are you subject to any new agency data guidelines? Please describe your approaches to or plans for data management. What kind of impact do you expect these guidelines to have for the domain? How do you think your data can contribute to shared data resources in your field or across fields or domains?

5. What are some of the recent or upcoming changes in data work that come to mind? How do you think your data can contribute to shared data resources in your field or across fields?

6. What kind of computational infrastructure exists at your site now? How do you use it?

II. Data Sharing & Data Repositories – I'd like to talk a bit about data sharing & data repositories.

7. How do you decide what data will be shared?

8. What are the preparations involved with data that will be shared?

9. What are some of the data sources you use in conjunction with your own data? How are they shared with you?

10. What kinds of public access to data play a role in your work? What changes have you noticed with respect to public access to your data? To the data of others?

11. What role do data repositories play in your work?

12. Please describe the kinds of data repositories with which you are familiar.

III. Closing Invitation – We've covered a number of topics.

13. Are there any other aspects of data work that you think would be interesting to tell me about?

- May I contact you if we need clarification?
- Could you suggest another individual who might be in talking about their data sharing and data repositories?

Interview Guide – Project View

- Study explanation: This study is about project data work and repository relations.
- Request signing of permission forms.
- Turn on recorder.

1. Let's start with you saying your name, title and position here at the university.

2. What kinds of projects do you work with?
-Tell me about the driver for project X; the data work for project X.
-What kind of data-related roles are there for the project?

3. Could you explain briefly your role with the project?
-How did you come to be work on this project?

4. Give me a brief description of how the project started.
-What kinds of data work are involved?
-What kinds of supports exist for the project? (e.g. digital, social, technical, organizational)

5. Describe some of the major decisions made or events for the project over time?

6. What kinds of data are generated and worked with? Are there standards?

7. What kinds of information infrastructure or computational arrangements – computers, instruments, technical assistance, and applications – exist for the project(s)?
-What kinds of relations are there with other research scientists, groups, partnerships?

8. What services do you provide other researchers?
-How do you envision the future of data work?
-What other researchers/groups/projects/networks do you find important to your work?

9. What terms are used to describe the aggregations of data you work with?

e.g. data systems, repositories, archives, facilities?

10. What roles are distinguished in those who work with data?
    e.g. data managers, data curators, and informaticians?

11. How has project X data work influenced your own work?

12. What have you learned about data work from project X?

* Closing invitation: We've covered a number of topics. Are there any other aspects of data work that you think would be interesting to tell me about?

- May I contact you if we need clarification?
- Could you suggest another individual who might be in talking about their data sharing and data repositories?

APPENDIX B. EcoPrairie Timeline and Project Data Work

B.1 EcoPrairie Timeline

1862 Colleges of Agriculture at land grant universities established by the Morrill Act

Period 1
1937 USDA Forest Service Central Plains Experimental Range (CPER) designated
1940 Grazing at CPER begins
1954 Pawnee National Grasslands transferred control to the Forest Service

Period 2
1968 ARS CPER begins working with Natural Resource Ecology Laboratory CSU
1969-1974 IBP/CPER: Rangeland Sustainability; Central Plains Ecological Research
IBP International Biological Program, Prairie Grass Biome (Risser)
1980 LTER proposal not funded

Period 3
1982 CPR LTER I proposal funded; LTER Central Plains Experimental Range site supported
1987 SGS LTER II proposal funded; Shortgrass Steppe site (CPR project renamed SGS)
1991 SGS LTER III proposal funded
1996 SGS LTER IV proposal funded
1996 SGS Using Department of Forestry, Sun Servers
1996 SGS Website established through university infrastructure

2000-2006 CSU Library participates in Colorado Digitization Program (CDP)
2000 CSU ISTeC Information Science & Technology Center established
2001-2009 CSU Library Digital Collections using ContentDM

2000 SGS data access provided online
2002 SGS LTER V proposal funded
2002 SGS Using Dept of Soil and Crop Sciences, Windows based system Ascalon
2002 SGS DM equivalent of a full time position with student help

- 2002-2007 period of rapid transition in LTER with data work

2005 CSU Task Force commissioned biennial Future Vision 2020 symposium
2007-2015 Colorado Alliance of Research Libraries, Alliance Digital Repository

- CSU Library a planning member only

2007 CSU Library obtained DigiTool partnership license with Ex Libris, implementation start
2008 CSU Institutional Repository begins
2009 CSU Geospatial Centroid begins at NREL

2008 SGS LTER VI proposal not funded
2008 SGS LTER VIb proposal probation 2 years

2008 CSU Libraries joined with Academic Computing and Network Services (ACNS)
2011 CSU Library Digital Collections of Colorado (DCC) partnership formed using DigiTool

2011 SGS LTER decommissioning funding

- 2 years to Feb1 2013 (plus 1 year no cost to Feb 2014)

Period 4
2012 SGS LTER discussions begin with Library re SGS data
2013 SGS IM envisions hybrid model for data migration
2013 CSU NREL Geospatial Centroid moves to library location
2013 SGS physical site at CPER closed
2014 SGS LTER site closed

2015 CSU Library Digital Collections software migrates from DigiTool to DSpace

B.2 EcoPrairie Project Data Work Expanded

This expanded description of EcoPrairie project data work includes six sections: the role of data management, project collective data work, relations with partners, closing data activities, working with two archives, and stories of participation.

An overview of the grant funding in six-year cycles is shown in Table B.1 that captures the change in the lead investigator together with the project's home department. While six lead investigator changes are shown for eight funding cycles. There are three data/information managers for this same time period.

Table B.1: A summary of EcoPrairie LTER funding cycles, 1982-2014

Org / Years	1982-1987	1987-1990	1991-1996	1996-2002	2002-2008	2008-2012	2012-2014
Cycle	**LTER I**	**LTER II**	**LTER III**	**LTER IV**	**LTER V**	**LTER VI**	**SGS LTER Sunset**
Lead PI	Woodmansee	Lauenroth	Lauenroth	Burke	Kelly	Antolin	Moore
Data Managers	Kirchner	Kirchner	Wasser	Wasser	Kaplan	Kaplan	Kaplan
Affiliation at CSU	Natural Resource Ecology Lab	Natural Resource Ecology Lab	Department of Range Science	Department of Forest Science, and then Department of Forest, Rangeland and Watershed Stewardship	Department of Soil and Crop Sciences	Biology Department	Natural Resource Ecology Lab and Department of Ecosystem Science and Sustainability

B.2.1 The role of data management

The LTER incorporation of data management as a site-based activity facilitated project data assembly and local data work. Embedding data expertise near to data generators provided help to individual project participants as well as support for project-wide and network-wide data activities.

Within the EcoPrairie project

Just as LTER ecological research efforts were anchored at the individual project or site level, so too were LTER data management efforts. From it's beginning, EcoPrairie supported a part-time data management position (see Appendix F.1 for a history of LTER information management). In addition, the project supported a field crew coordinator as well as a project-based field team at the field station. The first data manager was a researcher active in modeling as well as local and network data work (Kirchner et al., 1995). With digital technologies and Internet capabilities advancing rapidly, the role evolved at EcoPrairie combining data and field team responsibilities from 2002 to 2008. Information management efforts initially focused on expediting data entry while documentation efforts included development of metadata forms and workflow processes that were informed by metadata efforts at other LTER projects. At the start of the sixth funding cycle in 2008, the data position was changed into a full-time information management position that maintained ties with the field but no longer led field crew coordination responsibilities.

Data management efforts remained closely associated with activities in the field. In addition, as is typical in ecology, there was strong human and social engagement with nature and field activities (Roth and Bowen, 2001; Bowen and Roth, 2002) that differed from what Kingsland (2010) refers to as "the narrowing tendencies of laboratory work". The connection with nature was appreciated as a counterbalance to the digital environments within which contemporary data managers work. During EcoPrairie closing, a data manager who volunteered for fieldwork, mentioned a connection with the field:

> This was the first winter that I haven't had a rabbit count and as brutal as those winter rabbit counts are, I miss volunteering for them. That has been a sore spot for me. This is the first field season, there's no [EcoPrairie] crew going out there. (DM)

EcoPrairie data managers were often involved in collaborative efforts. For instance, team members consulted with telecommunications and information technology professionals on cyberinfrastructure improvements for a project-related Research and Interpretation Center. They cooperated with National Ecological Observatory Network (NEON) operations managers and field ecologists when NEON began working in this region. With a NEON instrumented tower established at CPER, EcoPrairie data workers consulted on site-specific techniques for experimental design and methodologies as well as on sample and data processing procedures developed at EcoPrairie in working with meteorological station data streams. Procedures involved download and parsing services together with QA/QC and formatting processes that supported submission of weather to the LTER all-site climate data system (Henshaw et al., 2006). In addition, EcoPrairie's information manager participated in the Grasslands Data Integration (GDI) cross-site database project (Kaplan et al 2007; Cushing et al., 2008; Vanderbilt et al., 2009). GDI brought together ecologists, information managers, and computer scientists to address the challenges of integrating long-term annual above ground net primary productivity (ANPP) datasets where data comparisons are difficult due to differences in methodologies and experimental designs, temporal and spatial scales as well as in species nomenclature and codes.

An information management team grew to include a half-time IT/Geographic Information Systems (GIS) manager, a part-time student web developer, and two quarter-time data entry students but by 2011 when the project was on probation, only two members remained: the information manager and the part-time IT/GIS manager. Both received decreasing LTER support

until the site closed in 2014. With support for the lead information manager at 25% toward project end, a variety of other projects within the laboratory began providing support for the work of data management:

> Every couple months I have to keep looking and seeing where the money is coming from…I've been doing database consultation, synthesizing data, building some websites, testing websites, working on sequel server implementing referential integrity and data quality, processing. It's like a hodgepodge of stuff. (DM)

Laboratory members during this time often asked for help in writing data management plans. One research activity within NREL envisioned a generalization of their existing information system for biological data, and another involved plans for coordination with institutional repository efforts. A road map was generated for a local data system model with a web-based data access and delivery system. In addition, the data manager continued to be involved in data management training for students and researchers as well as in publishing in scholarly venues (e.g. Wang et al., 2015).

Within the LTER community

A Science Council and an Executive Board coordinate scientific research across the LTER network of member sites while an Information Management Committee (IMC) coordinates data activities. The lead information manager from each project site is a member of the IMC, an LTER standing committee described as a community of practice (Karasti et al., 2006). The committee is framed by the LTER governance structure (Michener et al., 2011), IMC terms of reference (Baker et al., 2010), data policies (Porter and Callahan, 1994; Porter, 2010), and community metadata standards (Millerand and Bowker, 2009; Michener et al., 2011). LTER data work is documented in a series of guides and best practices (Appendix F). As a member of this network-wide IMC, the EcoPrairie lead information manager who participated in and co-led a number of network activities was immersed in a continuing LTER community discourse on data and information management.

The Long Term Ecological Research Information System (LTER NIS, nd) is a suite of applications that developed beginning in 2006 (Servilla et al., 2006; Michener et al., 2011) at the LTER Network Office (LTER Data Portal, nd). At the time of the EcoPrairie closing, NIS was transitioning from a development phase to an operational phase. An LTER Data Portal (nd) supported by NIS included a back-end service framework dubbed Provenance Aware Synthesis Tracking Architecture (PASTA) that managed data, metadata, and data packages (Servilla et al., 2006, 2016). A brief history of LTER NIS is provided in Appendix F.1. LTER NIS was developed after more than a decade of work developing, deploying, enacting, and revising a community metadata standard. With growing awareness of the importance of metadata quality and completeness, system developers partnered again with the LTER IMC to design a metadata validation module to check dataset metadata upon ingestion into the network system (Servilla and Brunt, 2011; Chamblee and O'Brien, 2013).

Metadata validation and data loading were major data management activities during the EcoPrairie closing period. The process for contributing updated metadata to the LTER NIS began as a project to establish a new online location for data harvest. After preparing a metadata file in XML format that follows LTER Ecological Metadata Language (EML) guidelines, the file is copied to a harvest location for upload to PASTA. This activates a metadata validator that assesses quality and correctness, assessing whether the data are 'PASTA-ready' (Bohm et al.,

2012; PASTA QC, nd). A harvest report is generated and sent to the data provider. If the data package is complete, registration occurs and a Digital Object Identifier is issued. Successful upload results in availability of the data in the LTER NIS data repository. NIS provides a searchable online interface and replication of data within the data aggregator DataOne (nd).

B.2.2 Project collective data work

Four topics associated with collective data work of projects are discussed below: local data management support, technical systems support, data assembly, and a project website. These topics captured the scope of data work in the subsequent two cases as well.

Local data management support: Collective practices

Refinement of data management processes occurred periodically and became a routine part of EcoPrairie project data practices. Tasks included inquiring about data availability, overseeing data entry, assembling data files on the project server, and gathering metadata. This project-related collective data work, was in turn made available to the researchers. For instance, a field manual developed by the EcoPrairie information manager was made available digitally online. The materials were also assembled in a field binder available to all project participants at the field station. The binder with printed pages was updated annually and became a central location for finding information about field studies and sampling procedures with ease of access for researchers and field crew alike. The binder displayed documentation incorporated in the project metadata forms:

> I've always had the PI really focus on our LTER field crew manual that is basically a word document. I've just been copying and pasting into the digital metadata file. And then I've been asking them for the missing pieces. Like the abstract. (DM)

In the manual, rules about safety, roads, and fieldwork preceded a listing of studies conducted at CPER. The list represents a hardcopy catalog of the year's field studies. Associated study-related information included objectives, study area location and design, sampling protocol, quality control instructions, and field data sheets.

Digital data files with accompanying metadata were assembled on a local project server by the EcoPrairie information manager. Measurements and observations from each field study not in digital form were entered into spreadsheets as data tables. Associated metadata was entered by project participants including students as well as the information manager who was responsible for keeping this task on track despite differences in individual participant timeframes. Three categories or versions were used to describe the phases of data processing: raw, working, and final. The project data server was used for assembly and storage of raw or relatively unprocessed data files from the field as well as for working and final data files that had undergone processing and analysis. These data files were organized using the server's hierarchical file management system with data organized by project name and by kind of data. Final metadata was input into a Microsoft Access relational database. The server also held GIS products. While there were some automated quality control procedures for climate, vegetation cover, and vegetation density, many data-related tasks involved manual procedures.

With the Agricultural Research Service (ARS) managing the CPER land as well as the research at this site, great care was taken by the LTER information manager with the ARS study number assigned to each sampling permit issued. Cross-referencing with this ARS unique

identifier was recognized as critical both to support collaborative ties and to maintain provenance of the data. ARS was an agency that did not have active systems online so EcoPrairie developed an online system for assigning permit numbers to each project granted permission by ARS to sample at CPER. With EcoPrairie closing, a return to downloading and mailing PDF permit forms appeared likely.

Local systems support: Project server

With the digital era well underway by EcoPrairie's fifth funding cycle (2002-2008; Appendix Table B.2.1), the affiliated department supported a local systems administrator that facilitated purchase and maintenance of a server for EcoPrairie. This technical support continued until the project ended. A staff member from this department was part of the EcoPrairie information management team, providing systems and applications support as well as leading geospatial data activities.

After the fifth funding cycle rotation of project leadership, however, the project and data management offices moved back to NREL for the sixth or closing cycle (Table B.2.1). NREL maintained its own systems management group that consisted of two specialists performing systems administration, consulting on system services, and interfacing with university-wide IT services. With EcoPrairie technical support for the server remaining in the previous cycle's department, this meant first that the data management office resided at a distance from its technical base of support, and second, that the EcoPrairie data management consulted with but was not tied to NREL technical services. This split of the social and coordination of data management from the technical aspects including system administration was not considered a major difficulty given research in co-investigators resided in both departments.

A single project-based data system

Two key elements of a digital data system are assembly of data from multiple sources and online access to the data (Michener, 1986). Both elements are receiving increasing attention as familiarity with the concept of a data repository grows (e.g. Ray 2014; Johnston, 2016). In speaking about the assembly of their data into a local repository, EcoPrairie participants continued over time to refer to their 'data system' or 'information system', terms dating from the site's first decades of aggregating data. EcoPrairie researchers in time became familiar with project data practices and final products though not necessarily with the particulars of how data management was accomplished. One EcoPrairie researcher referred to the usefulness of local data arrangements:

> So we went back … and got the data we needed. I mean we had those data ourselves, but there was something about it … some part of that dataset that, somehow, the way it was stored in LTER, that made their dataset useful. (RS)

A data file containing measurement and observation data in matrix format has columns with variable names that were defined in a specific manner at EcoPrairie. Dating from early efforts of their first data manager, an EcoPrairie data practice developed for each data file a listing of the variable names it uses in a file called a '_var table' where each row or record in the table describes a variable. The _var table has columns with the variable's name, description, measurement type, ratio interval, and unit of measurement, all information needed for complete EML metadata. Due to the simplicity and structure as well as the presence of a data management

advocate for the practice of using this named entity, the '_var table' approach to documentation became a local convention familiar to project participants.

Project-specific data assembly was carried out largely at the sites during the third decade of the LTER. Metadata including the variable information for each dataset at EcoPrairie was added to a Microsoft Access database along with lab protocols, standardized Geographic Positioning System (GPS) data, and use of controlled vocabularies for keywords. Units and variables were coordinated with an LTER community unit registry. When funds were made available that leveraged previous work including an all-site assembly of datasets by the EcoTrends Project that began in 2004-2005 (Peters et al., 2013), a centralized LTER Network Information System (NIS, nd) was envisioned (Servilla et al., 2006). By 2007 Perl scripts together with an XSLT (Extensible Style sheet Language Transformation) at EcoPrairie were generating EML files in XML format from the project relational database and storing them on the web server for harvest by the LTER NIS. This level of metadata delivery to a centralized system facilitated network-wide data discovery and access. EcoPrairie datasets together with other project data continued for the duration of the project to be made available on a local project website that was established prior to work with NIS.

Website for a single project

An EcoPrairie digital presence was established via project web pages that delivered news and overviews of the project, the research, and the education/outreach efforts. In addition to a bibliography of project-related publications and reports, data was made available. Web content grew in time to include summaries of project topics and a data catalog containing both datasets and maps. A website redesign in 2008 was carried out in a participatory manner reaching out for input from staff, researchers, and students as well as to Creative Services, a campus communication office. By 2010 a dynamic data catalog was online in addition to new mapping tools for viewing and downloading spatial data layers via an Internet browser.

The EcoPrairie data manager prepared and managed content for the project website using the campus Academic Computing and Networking Services (ACNS) unit web services. Web pages were created in HTML. An information system developed using content management software called Drupal Ecological Information Management System (DEIMS) (Aguilar et al., 2010; San Gil et al., 2010) was explored. DEIMS is an application designed and used by a number of LTER and ILTER network members to manage scientific information - data, metadata, catalogs, and directories – with web delivery. It was decided at EcoPrairie not to pursue the DEIMS option because the university-wide support for EcoPrairie's website had expertise with Microsoft applications but not with Drupal.

B.2.3 Relations with partners: multi-sector partnership

A long view of partnership at Colorado State University (CSU) dates to its history as one of more than 70 United States public land-grant colleges created beginning in 1862 (NRC, 1995). Land-grant colleges were established in order to bring together agricultural and technical higher education with research and to ensure 'extension' outside academic circles. Thus partnership and communication among academic, government, business, and professional associations were established as a central tenet for CPER at the outset. Academic research funding supported the IBP and LTER programs (Periods 2 and 3) during which time USDA/ARS worked synergistically with these programs. During these periods, ARS benefited from close academic

relations especially given the access to modeling efforts tied to field data collection, to graduate students exposed to ecosystem thinking, and to experience with collective data management.

As the CPER site-based LTER program ended, the Long Term Agro-ecosystem Research (LTAR) program began. Launched by USDA in 2011, LTAR was planned as a 30-year national program in water availability and watershed management for agriculture (LTAR, nd; Robertson et al., 2008). Drawing on the network of land grant colleges together with the network of geographically dispersed agricultural experimental stations (USDA/ARS), USDA initiated an eighteen-member network comprised of both ARS and non-ARS project sites. The program includes coordinating data efforts via the National Agricultural Library, one of only a few national libraries in the United States (U.S. Libraries, nd; Beachy, 2010). Among the first sites funded was the ARS led by the Rangeland Resources Research Unit located at Cheyenne Wyoming, at Nunn in Colorado and at the CPER site at Fort Collins.

With forested and agricultural settings, the LTAR experience aims to address land-related grand challenges. Plans include coordinating field measurements via standardized methodologies and protocols together with shared research questions. In addition to the shared-theme approach to consideration of LTER, the LTAR approach aimed to coordinate field methods at the outset. Further, each site is addressing data management individually as the network adopts a three strategy approach that includes: a systems approach to ecological studies, a network approach with site-based research project members, and a progressive approach to building a 21st century workforce (e.g. NRC, 2015; Varvel et al., 2011; Swan and Brown, 2008). The program draws on experience of earlier programs such as LTER, NEON, and the National Institute of Health partnership with the National Library of Health Sciences. An LTAR researcher explains the program by analogy:

> LTAR is like NEON but not quite as strict as NEON. It's like LTER but more top down. Its almost like it's somewhere between NEON and LTER. (RS)

One discerns the juggling of lessons learned from previous programs in this exceedingly brief, informal description of plans for a new network approach. Support for a data management position is explicit in LTAR planning documents, thereby bringing some measure of continuity to data holdings in the transition from LTER to LTAR.

With the long-standing partnerships across laboratories, departments, institutions, and sectors at this place-based research location, there is a critical mass of scholarly research and a community tradition of applied use of the knowledge regardless of succession in leadership and funding. Work at CPER included not only leadership changes within the established LTER program (Period 3) but also leadership moves between academic and government organizations during Periods 1 to 2 and Periods 3 to 4. This case study in focusing on data work does not address the policy and management science issues associated with termination nor does it consider issues of succession except to note that partnering anchored at CPER over the years since 1939 provided continuity to scientific research of the shortgrasse steppe biome and of grazing studies. The cadre of researchers from across the sectors was able to continue field-based research at CPER due to its status as an experimental site as well as its diversity of potential funding sources. The continuation of sampling under the auspices of a new program initially appears fortuitous though its century long history suggests some advantage to multi-sector, multi-stakeholder partnerships focused on sampling at a defined location.

The LTAR in Period 4 represents an opportunity to observe another network initiated with data managers embedded at each site and a centralized information system in development as a standardizing presence launched during network formation rather than two decades after the start of the network as occurred for the LTER network. How will data work be distributed? What timeframes will be needed to establish submission of data to the network system(s)? What will be the distribution of data work between local sites and a centralized network system?

B.2.4 Closing data activities

The closing date for EcoPrairie was February 1, 2014. A final two-year supplement (2012-2014) outlined four information management goals:

- Goal 1 - Improve Data Access: Complete the delivery of Level 5 EML 2.1.0 compliant metadata and data of core datasets as data packages through the LTER Data Portal at the LTER Network Office while satisfying existing best practices and standards for the LTER Network.
- Goal 2 - Preservation of Data Legacy: Augment the data packages developed through Goal 1 with additional supporting documentation that reflects the legacy of the site and studies.
- Goal 3 - PASTA Compliance: Ensure that [project] data packages are interoperable with the network infrastructure and the emerging PASTA Framework.
- Goal 4 - Network Information System (NIS) Participation: Extend involvement in the LTER NIS development initiatives by participating in the broader information management community as a member

To achieve these goals it was necessary to ensure packaging and upload of project datasets to the LTER NIS system (Goal 1) with data that met metadata and data requirements (Goal 3). Goal 2 was described in the supplement as "a unique opportunity to embody innovative pathways to ensure accessibility of data, information, samples, and other products associated with the [project] legacy". The general language of this goal provided leeway to be able to respond to contingencies at a time when well-established processes for data preservation in repositories were unsettled and when distributed data access for heterogeneous data was not yet well understood. In this case, response efforts included investigating how to provide preservation of supporting documentation and other project materials as a whole, how to create repository-to-repository connectivity, and how to continue time-series data beyond a snapshot in a membership repository for projects that cease to be members of the network. Finally, Goal 4 identified a need for support in order to continue participation in network-wide discussions about updates in the EML metadata standard and the NIS metadata validation developments. It also represents a statement, regardless of perceived awkwardness or reluctance, of the need to interface with a terminated project. The EcoPrairie information manager recognized there were insights to be contributed from projects undergoing termination. Lessons learned about data migration were captured in a report and a list of data-related project closing activities included critical procedures for members leaving a network. See the third and fourth stories of participation in Appendix B.2.6.

Closing activities for the EcoPrairie information management team included consideration of how past research projects brought their efforts to a close. Three LTER project sites closed in the first two decades of the network. Their closing activities, however, were little

documented and occurred prior to when the Internet made websites and online data systems possible. As a result, names of personnel from the closed projects appear in an early hardcopy LTER-wide directory but do not appear in the subsequent online network-wide personnel directory. Whether EcoPrairie personnel will remain in the network's personnel directory is unknown, but it was decided by the project information manager that it would be sufficient to assume that the names of EcoPrairie personnel would be available via EcoPrairie online materials such as datasets, proposals, and reports. It was also decided that finalizing a project bibliography would be difficult and a source of irritation to authors of publications left off such a list so the idea of a project-generated bibliography was abandoned, deferring to new alternatives such as the Web of Science and Google Scholar. Thoughts on preserving the project website changed a number of times. Since much of the website content was captured as digital objects in a project collection and support did not yet exist for website archive at the university, ultimately no other plans were made to preserve the project website beyond what is captured by The Wayback Machine (nd).

The EcoPrairie database with active server pages made possible online search and query of the dataset metadata as features of the project data system. This capability was removed during the closing cycle when the university IT informed EcoPrairie their technology was outdated and no longer supported:

> And so we have to get rid of the dynamic piece of it [the data system]. And what we are going to be doing is only serving up static information. But where that static information sits is in question. Should it sit within the laboratory website? Should it sit with the Shortgrass Steppe Research and Interpretation Center website? Should it sit with the collection at the library? (DM)

"The dynamic piece of it" referred to the linking up of a local data server to a networked web delivery system for content delivery. At closing, a copy of the database remained on the project server but did not interface with the web, thereby minimizing maintenance and ensuring availability of content locally for report generation and other post project activities.

One researcher referred to the data management effort as making a continuing contribution and doing 'really well' during the project termination period. This researcher expressed what was repeated by a number of the project researchers interviewed. Implicit in such a statement is that the work of information management was not only appreciated but also recognized and accepted as an element of contemporary, collaborative scientific research.

The EcoPrairie data system was a project-specific element of data infrastructure. The shutdown of the data system and project data management activities was assumed rather than explicit in the project termination. The final two-year project supplement, however, provided critical time needed for making closing arrangements. The full functionality of the system was abandoned rather than modified or adapted for use with other projects because the technology was dated and data handling was largely manual. The server remained with static files and was accessible to project data team members in the know.

The EcoPrairie final project office was in a laboratory active in designing information systems with support from local systems administration staff and researchers but the design and purpose of their systems differed from those of the EcoPrairie data system. With no funds available for update or redesign of the EcoPrairie system, the information manager began looking to the future while commenting on loss of the project data system:

> It's going to be harder to leave here than I thought. … Like do you need an institutional repository or a project repository to work with? Do you need the server? Or maybe not. Maybe we're going to a different model or maybe different clients have the infrastructure to work with. (DM)

That is, project data infrastructure was perceived as gone, and there was uncertainty about a new set of infrastructural elements and how they might support data work that is increasingly digital, collaborative, and public.

EcoPrairie's lead information manager at closing, a local research assistant familiar with the ecological domain who had acquired data management skills on the job, was faced with finding a new position. Within EcoPrairie's soft-money funded laboratory NREL, advancement in data-related positions was not seen as contingent on a degree but on a demonstration of leadership through success in funding. At the same time, the campus library and the USDA ARS posted job advertisements for new data management positions (Appendix G). These were of interest to the EcoPrairie information manager who considered the position with the library:

> I have this vision in my head of trying to get from a central location like the library to provide, basically I can imagine myself as a liaison between the scientists and the library for research data for our lab or our college with researchers in the domains of data I know… And then, helping them with education. This semester I've done four training sessions for different venues of people in the lab. (DM)

The position for a data management specialist in the library was initiated in a library track requiring a professional degree in Library and Information Science adjunct status was offered to those without a library degree (Appendix G.1). After considering the strengths of a library position that provided both training and continuing education as well as the limitations of an adjunct position within academia, the EcoPrairie information manager applied for and accepted a position advertised by the USDA ARS so remained with the CPER sampling site (Appendix G.2). This job ad described a position that "serves as a support scientist responsible for the data management activities". The position was offered as a category three scientist, a position where advancement neither required nor precluded scholarly publishing. It was supported as part of the new LTAR program. The EcoPrairie information manager stressed that in addition to working within a collaborative network and preserving the data, with LTAR "there seems a priority to serve the dignity of the PI" as well as availability of training. Further, it was noted that national libraries have lines of funding that differ from those in academia.

B.2.5 Working with two archives

The following section reports on an expansion of EcoPrairie's approach to archiving data. Initially work focused on dataset submission to the LTER NIS to the exclusion of other project artifacts. From a partnership with the library involving digitization of old reports, there emerged a realization that the EcoPrairie datasets and related artifacts could be assembled and presented together as a collection. The extended decommissioning period was key both to making updates required for submission of data to the LTER NIS and to exploring a new approach to preservation. In pursuing a second approach to archiving that differed from and complemented the LTER strategy, the concept of a hybrid model emerged.

Working with the LTER NIS: Project datasets

At closing, there were a total of 105 EcoPrairie datasets with validated metadata archived in the LTER NIS system. This was achieved despite three issues that arose: updates in metadata validation criteria, incompatibilities in database constraints, and developments in dataset citations. The need for complete metadata, required for automated systems to perform as designed, added unanticipated work to EcoPrairie closing activities. Community metadata compliance criteria updates, initially scheduled to occur after the EcoPrairie project closed, were launched instead early in January 2013 in response to funding agency re-scoping (Chamblee and O'Brien, 2013). The EcoPrairie information management team described this situation as 'raising the bar' in the midst of their closing. The original metadata validator designed with three checks expanded in stages to five, then fourteen, twenty-seven, and eventually thirty-four completeness criteria. This required the upgrade and resubmission of the project metadata files. Supplements requesting support for updating project metadata to meet the EML 2.0.1 standard and then EML 2.1.0 were submitted by many LTER projects including EcoPrairie. In April 2013 EcoPrairie succeeded in uploading its first updated metadata file to the NIS PASTA system. The ANPP (Above Ground Annual Net Primary Production) file, a time series running from 1983 to 2012, was a core dataset described as "all integrated together and cleaned, not a controversial dataset". By May 2013, it met the twenty-seven checks required at that time. Since the 'outdated' EcoPrairie server was offline, another month passed before access to datasets was staged at the institutional repository so that the data files described by the metadata could be successfully harvested from a distribution URL included in the metadata. By midyear 2013, the EcoPrairie data submission process and the LTER NIS ingestion process were aligned.

Metadata issues arose again in December 2013, however, when data package warning errors appeared. Investigation revealed the issue was due to differences between acceptable structuring for the local and remote database systems. Changes to variable names happen over time during fieldwork activities. This led to different ordering of variables in the database _var table and the table of data measurements. An EcoPrairie Microsoft Access database was able to match up variable names regardless of order but the NIS PASTA system required the ordering to be the same. Although only a report warning was issued, not a fatal error, the EcoPrairie information management team changed its metadata arrangements since other systems in the future might have these more stringent requirements. Though time was short, EcoPrairie assumed responsibility for fixing not only fatal errors but also warnings:

> I keep thinking we should take the high road, we should do the best we can, we should do this … We are going to give it a shot. … There are going to be bugs, and there are going to be errors, and we are going to need help. (DM)

Once variables for all the datasets were standardized to match the order of the variable columns in the data files, NIS reports on metadata submission and data harvesting were again free of warnings.

Another issue that arose to make submission of project data difficult during the closing period related to information about the project included in the metadata. Project closing raised awareness of the need for greater clarity in describing the project in the dataset title, the project description, and the dataset citation. New local conventions were developed to ensure a clear description of the project appeared in data catalogs. Citations often include four metadata

elements (title, publisher, date, and dataset creator). EcoPrairie added two items to the title: name of the project and the project's sampling permit number. The following is an example of a title:

> [EcoPrairie] Bouteloua gracilis Removal Experiment Vegetation Density Data (ARS #155) on the Central Plains Experimental Range, Nunn, Colorado, USA

When searching then for 'Boutelou gracilous', the new title conveyed that the dataset is associated with both EcoPrairie and ARS. Project information and links were also added to the metadata abstract. The EML abstract content was expanded by preceding the dataset particulars with a paragraph about the project and the project collection:

> This data package was produced by researchers working on [EcoPrairie] Project, administered at Colorado State University. Long-term datasets and background information (proposals, reports, photographs, etc.) on the [EcoPrairie] project are contained in a comprehensive project collection within the Digital Collections of Colorado (http://hdl.handle.net/10217/100254). The data table and associated metadata document, which is generated in Ecological Metadata Language, may be available through other repositories serving the ecological research community and represent components of the larger [EcoPrairie] project collection.

Finally, the issue of dataset citation was developing rapidly after 2010 in conjunction with data sharing guidelines in community forums (Goldstein et al., 2017; Starr et al., 2015; Duerr et al., 2011) and in practice within organizational settings (e.g. Mayernik et al. 2012). NIS created an LTER citation from the EML metadata file that contains creators, title, date, and publisher. Upon submitting this information to EZID (nd), a DOI issued by DataCite (nd) was assigned and could be added to its presentation online. An example of an online citation for the Bouteloua gracilis dataset included dataset specifics as well as a study number and a network name:

> Lauenroth, William (2013): [EcoPrairie] Bouteloua gracilis Removal Experiment Vegetation Point of Intercept (Cover) Data on the Central Plains Experimental Range, Nunn, Colorado USA 1997-2005, ARS Study Number 155. Long Term Ecological Research Network. http://dx.doi.org/10.6073/pasta/973bc5c24e01e375b42abf7ac9774447

Working with the library: Project collection

Though project termination was a difficult time, having the project end in sight and a dialogue with the library ongoing, prompted thinking about what constituted a 'completed project' in terms of EcoPrairie's legacy. The availability of digitizing services at the library and the capability to assemble project artifacts in turn spurred thinking at EcoPrairie about the concept of a variety of digital products. The issue of handling 'other products' took on urgency for EcoPrairie with the flood of materials flowing out of offices and file cabinets that were emptied into the project administrative and data management offices in anticipation of project-end reallocations of space. EcoPrairie learned about the library history of work with 'finished' collections as final products and began work with the library on an EcoPrairie project collection. Strong mutual interest was identified relating to collection formation and the library's interest in expanding its repository to include research data. The library brought to the partnership ties with

the university-wide Academic Computing and Networking Services (ACNS) unit that resided administratively within the library at this campus. That is, information technology services were organizationally connected through the library to address data management issues such as networking, storage, archive, and access that were under discussion at the campus-wide Information Science and Technology Center (ISTeC). Three data-related activities carried out with the library involved digitization, data access, and data migration.

The first overarching data activity in partnership with the library involved digitization working groups focusing first on project reports and subsequently on project photographs. Digitization efforts began in 2006 with an EcoPrairie decision to move historic reports from the 1930s to the campus library. These hardcopy reports resided on shelves at the laboratory as well as at the field station. A few times each year, there was a request to the EcoPrairie information team for material in one of the reports. Copies were made and mailed or scanned and sent electronically as a pdf. Print reports, however, were at risk of being lost as they were subject to borrowing and to deteriorating conditions due to insects and dust. When the EcoPrairie information manager discovered the library digitization capabilities, the IBP reports were digitized and made available through the library digital collections (IBP Reports, nd).

The photograph digitization working group drew on the library archive's experience with digital photographs (Hunter et al., 2010) and played a major role in preservation of EcoPrairie's old prints and digital photos. EcoPrairie's loss of project space upon closing led to discovery of a variety of kinds of materials and a need to work with staff associated with the print archives unit that in turn worked closely with members of the digital repository unit:

> And out of all of that material that emerged, the photographs took on a life of their own because we were working with both the traditional archive as well as the digital archive. (DM)

Motivated by finding aides for a water digital collection in the Institutional Repository, EcoPrairie digitization efforts ultimately brought together participants from multiple library units including ACNS/IT supporting collection software, a Special Collections archivist managing physical materials, and a digital collections librarian. Students trained by archivists carried out scanning. Digital files were initially stored on EcoPrairie's local project server with their other digital materials. The information manager engaged an EcoPrairie staff member with time available and an interest in photography to both reformat and create metadata for each of the photographs found in project and individual researcher files. A metadata template was developed in conjunction with the library. For digital photos, basic criteria were developed since photos ranged from thumbnails of 15 kilobytes that are typically discarded to photos of more than a megabyte that are considered for archive. Digital image data handling was documented in procedural guidelines to address CSU Library Metadata and Selection Criteria for Digital Images (Kaplan et al., 2014b, Appendix 10.5). The final outcome was a series of digital photographs placed as a collection in the library digital collections that subsequently was linked to the EcoPrairie project collection.

A second overarching data activity relating to data access involved discussions with department researchers, data managers, librarians and colleagues at other universities. A professional curiosity motivated this inquiry into what was available across the data landscape given the various efforts ranging in scope from domain-specific data initiatives within the natural sciences to institutionally specific endeavors within the library sciences as well as laboratory

specific endeavors such as the International Biological Information System (IBIS, Crall et al., 2010) within NREL. Change was evident as the Geospatial Centroid project center (Geospatial Centroid Center, nd) was moved from its development at NREL to provide GIS services within the library. One NREL researcher articulated a view of data work as a field of inquiry:

> the IT/IM component when I started … the IT was the technical side of machines and hardware and capabilities. We have an IT staff. IM was largely the purview of the individual projects. … now the whole nature of the relationship between data and science is changing. We call it big data now. … think about it not in terms of an information management service to projects but about the scholarship behind it. Think about what the data architectures look like. What are questions that you could ask that you could only glean from having large datasets? There is a dynamic to the data itself that we are missing when we just study little bits and pieces in isolation. (RS)

At NREL, where the EcoPrairie project office resided, the laboratory's digital infrastructure followed a 'pay-to-play' approach. EcoPrairie, however, had no funds or interest during its closing years to replace outdated project technology and techniques. EcoPrairie's data work was not supported as a research endeavor.

For the EcoPrairie information management team, the campus library with its growing number of digital curators and technologists represented a new partner with interest in managing data and potentially in working with project data managers on design of workflows and systems informed by local project data work. After 2000 within the United States library community, interest in development of institutional repositories (Lynch and Lippincott, 2005; Erway, 2013) together with management software contributed to growth of research data services (Lynch, 2008; Tenopir et al., 2012). The library state of transition is made visible by changing names for individual staff positions as well as for its subunits. A 'data librarian' was hired at CSU libraries to work together with their 'metadata librarian' during the EcoPrairie closing period.
The library represented another option in addition to network and domain archives for those engaged in EcoPrairie data:

> It's not that they [the LTER Network Office] have the only system. And it's not like they have this mainframe and we're still working with pen and paper here. The technology is ubiquitous so the centralized network technology is not the only solution available to you. So we have choices. (DM)

The library became an on-campus partner for EcoPrairie with experience with scholarly resources, an organized professional workforce, and a growing digital capacity and a specific interest in developing data services.

The third overarching data activity in partnering with the library was the formation of a data working group by the EcoPrairie information manager in 2013 (Appendix B.2.6, Stories 2 and 3) to explore data migration and support for an EcoPrairie collection. At the time, the campus library was using the digital asset management system DigiTool (nd), commercial software from Ex Libris for managing digital collections and institutional repositories. As a cost-saving measure, the campus library administered an Ex Libris DigiTool license for a consortium partners known as Digital Collections of Colorado. DigiTool, a relatively mature product with technical support available, was seen at that time as requiring less investment from local IT staff

than other digital asset management applications such as Fedora or DSpace. Further, DigiTool handled both simple and complex objects using established file formats and metadata standards. The software featured unique internal identifiers or handles in addition to providing simple web interface customization and usage statistics availability. Descriptive metadata options incorporated in the application included Dublin Core, MARC, and MODS, although the campus library primarily used a qualified Dublin Core defined by a Core Data Dictionary (Rettig et al., 2008; Hunter et al. 2008).

The working partnership between a research project data manager and the library was described as follows:

> It's sort of a win-win situation because they haven't dealt with research data. And I've got research data that needs to be dealt with. So its sort of like we are figuring out as it goes, as we go. And we're bringing in lots of other issues, both with a vision for the future and establishing a baseline … and adding new value to the datasets within a re-use scenario. (DM)

Bi-weekly meetings were held by the four-member data working group designated a library pilot project. The mission was to plan and develop a process for moving data from the EcoPrairie project server to the library digital repository. In anticipation of the EcoPrairie data system shut down, the concept of 'data migration' emerged to describe moving the data and associated materials to the campus digital repository. Group members included the EcoPrairie lead information manager, two data librarians, and myself. The group began with an exchange of information about practices, controlled vocabularies, and ongoing efforts followed by an inventory of EcoPrairie project digital and physical materials. Digitization efforts grew to include scanning of artifacts such as field datasheets in addition to organization of Geographic Information System (GIS) layers of topographic features, landmarks, and study sites. Over time, the project collection planning became inclusive of the data as well as the metadata in the interest of collection completeness though datasets were a new category of digital object for the library.

A first dataset for ingestion by a library development server was selected to begin the collection formation process. A dataset was chosen for transfer that included a suite of associated materials such as a thesis, published articles, photographs of plots, and spatial data. The concept of a primary ReadmeFirst file when curating at a collection level was discussed in conjunction with the collection of files presented as a zipped folder with data and metadata. A shared staging area was established in Dropbox (nd) to facilitate the movement of data from the project server to the library system since DigiTool, the library application in use, lacked a front-end ingestion interface. Using Dropbox for data provided a space to set read and write permissions as needed and to develop a structured set of folders that would be intuitive for participants. This transfer process made visible what became an ongoing grouping activity. The shared space enabled joint development of categories that in turn prompted discussion about data groups and names. A number of collection series were established over time: GIS layers, pasture treatment maps, photographs, presentations, progress reports, proposals, site reviews, species lists, and symposia. Digitized versions of the field manuals were placed online for the years 2007 to 2012. The data working group developed workflows for the categories or groups of digital files. Having project theses and dissertations available in the institutional repository and photos in digital collections with internally generated identifiers meant that data could be associated via links with past

scholarly work as well as with old photos such as those showing bison grazing and cowboys working on the prairie.

The data working group collaborated closely on metadata documentation and data access. Issues of a generic scope like augmenting existing library metadata required justification since library conventions did not exist for scientific research project collections. Questions sometimes went to the dean or assistant dean of the libraries. EML metadata tags used by the project were identified and mapped to the DigiTool basic Dublin Core metadata template in use at the library institutional repository using an XSL transformation to construct the map and generate other pertinent metadata elements prior to ingestion into the library digital collection (Kaplan et al., 2014b, Figure 7 and Appendix 10.8). Noting that only a subset of EML tags could be mapped to Dublin Core, an EcoPrairie data manager suggested the addition of EML tags to Dublin Core as follows:

> EML goes beyond discovery of traditional Dublin core since they had envisioned using it for research data. I added keywords, abstract, award# from EML to the basic metadata. I added the spatial extent, description of place, and something on rights. (DM)

This expansion of existing library practices with basic metadata to incorporate key elements found in practice was a significant step toward more complete research data documentation. The collection formation effort informed institutional repository developers about data issues associated with research project collections.

Conceptualizing a hybrid model

While improving metadata for LTER data package submission and developing a research project collection of datasets and other project artifacts in partnership with the library, EcoPrairie data management planning expanded to consider relations between systems. In working with a number of communities and kinds of systems, an understanding developed of data management as not just a quest to establish a single solution that met data needs but also for sending data to more than one archive and planning access to distributed application systems. For example, data in the EcoPriaire data system eventually also resided in a domain repository (datasets in the LTER NIS) and in an institutional repository (a project collection in the CSU library repository). Further, the project collection at the library provided access to datasets for harvest into the LTER NIS. Future planning involved considering how to provide access to application systems such as those specializing in visualization and/or synthesis (e.g. the Geospatial Centroid or GIS systems). Access to data was imagined as an interconnected ecosystem of repositories:

> I mean the connections within that web of repositories, I don't think has been thought out. (DM)

Interest in making data available to more than one repository required recognition that digital systems could have differing organizing strategies and content as well as different aims and capabilities. In developing a networked view of the data landscape in 2013, an EcoPrairie participant described a hybrid model:

> So it's a hybrid model where the library will house the data with this persistent URL so NIS and PASTA can go and get it, but other systems can go and get it too. (DM)

In time, the hybrid model represented a change as explained by a data manager, "We can do better than general solutions in terms of facilitating data re-usability by ecologists" and continued:

> So I'm trying to kind of lay things out in a way that the library does what they are good at. … Let them store it. Let them curate it. Let them make sure that it's available, you know forever, and that it gets migrated as technology changes …. And lets just access it from our local systems that really want to be doing something with the data. Doing some type of analysis, some type of forecasting, visualizing, data mapping. [DM]

The response of EcoPrairie researchers to the hybrid concept of a suite of applications varied. One researcher was interested in expanding their own data system at NREL to work with other kinds of data that would involve data management services for department members. This investigator recognized the expertise required for archiving data and realized their systems could map and visualize data in addition to synthesizing and summarizing data.

Several issues arose in pursuing the hybrid model. First, establishing machine access to datasets within a collection for automatic harvesting by other systems proved to be a need with which library partners were not familiar. Issues of access arose in using Digitool as a platform from which LTER NIS could harvest the EcoPrairie data files. To provide data access to datasets requires systems to have shared specifications. Currently, data files, data packages, landing pages, and collections have many different formats. When a data object has an assigned persistent URL (a local handle or a DOI), however, it is possible to retrieve it. Initial plans with Digitool were to bundle the EcoPrairie project collection files into a single zipped file for ease of download. Special arrangements had to be made so that data files would be machine accessible outside the compressed file and thereby machine accessible.

> The data package that goes into the Network Information System with the metadata that is PASTA compliant, has a distribution tag with a URL in it that gives you direct access to download the data. So if the datasets sit within a zipped file as a data package within the Institutional Repository then another computer could not just download the data. It would have to get to the zip file and then open it up and then find the actual data table in there. So we explored different options for interoperability or computer-to-computer access, direct access to the data table. … We thought about having another copy of the data tables somewhere else but redundancy is not good practice. … So we settled on bringing the two data tables outside the zipped package. (DM)

Initially there was professional concern about metadata separated from data files:

> Well somebody is just going to come and download data tables and they are never going to read the metadata … you know the attribute level metadata like the field and column header definitions that are necessary for a scientist to re-use the data. The only metadata they would see was what we were presenting as the Dublin Core fields that isn't detailed. But then I decided that I can't control everybody. At some point they [data re-users] have to take responsibility for re-use of the data. So we are just going to pull these two data tables out. (DM)

Here the information manager's desire to guide data users to read the metadata became a second priority in order to establish data system interoperability. Including the data both within the zip file and outside the zip file was considered as an option but the option was dismissed since redundancy introduces its own set of problems relating to duplication of files.

Staging the data collection in the library repository to provide machine data access raised a question regarding the relative standing of repositories:

> … mostly what defines it [the hybrid model] for me is the interoperability, that we are able to get between the two repositories … between the secondary repository and the primary repository. [DM]

An issue of status and relations among repositories arose because of the metadata developed but rarely used to date in describing registered digital objects assigned a unique identifier. Much like the case of ISBNs that assign unique identifiers to books, international agreements have been reached that establish DataCite, an organization that issues and manages Digital Object Identifiers (DOIs) (Brase et al., 2015). DOIs for data packages provide the information needed for data citation. The DataCite metadata schema includes descriptive elements that structure reference to multiple repositories related to the data. Anchored for years as a member of the LTER community and involved in work with the NIS, the lead information manager initially considered the LTER repository as the primary repository for EcoPrairie data. The EcoPrairie closing grant detailed the requirement for data to be in the LTER NIS, an indication that NSF also considered NIS to be the primary repository. Another perspective emerged as the library repository project collection took form and led to an alternative perspective of the project collection at CSU as the primary repository since it was a partner closer to the data origin that included a larger number of items in the collection. Discussions with staff at both repositories indicated neither was ready to address this issue, making it another aspect of infrastructure growth that was not yet in place.

Partnering of the EcoPrairie information manager with the university library began because of the library's development of digitization and preservation services. Coincidentally, a U.S. economic downturn during this time prompted discussions about the potentially greater stability of an institutional repository due to its place within library hierarchies and its position within the university:

> In 2011 there were a lot of government shut downs and I started getting concerned about [data] preservation. And that's when I thought, the institutional repository has its mission of preservation, and because of that may end up outliving repositories that are associated with specific research programs funded on soft money … I wonder if an institutional repository is somewhat protected from decommissioning? (DM)

The question was posed by the data manager "Who is most likely to be left standing in the end?" given the many kinds of data repositories – both long-standing and nascent. Repositories are a growing part of the research infrastructure supporting the sciences (Baker and Duerr 2016a, 2016b; Pampel et al., 2013), though their sustainability is an unresolved issue at national and global scales as well as at local scales (Berman and Cerf, 2013; ICPSR, 2013).

B.2.6 EcoPrairie stories of participation

The extended closing period of the EcoPrairie project provided time for developing data-related concepts and products detailed below (see Table 3.2). The following four stories describe my participation at EcoPrairie.

Participation story 1. Developing a hybrid model

The hybrid approach represented a shift from single-repository thinking when partnering with the library led to the formation in 2013 to the data migration working group by the EcoPrairie information manager. I was invited to join as a participant at its launch. The working group shared the experience of addressing differences in data management and data curation perspectives. Within the working group, perceptions shifted from seeing the selection of a single, self-contained data preservation strategy as an intractable assessment problem to seeing alternative options offered by a distributed network.

> You called it doing a hybrid model. I hadn't really thought about that. I kept thinking about it [the data] either going here or there. But I didn't even realize that what we were doing was like this hybrid model or that it really is situated within this vision of a web of repositories. Thinking of the big picture stuff gave me the ability to kind of think about and articulate it in a different way. (DM)

Giving the concept a name of 'hybrid model' was a way of acknowledging the existence of more than one potentially complementarity approach to data arrangements.

Participation story 2. A data migration poster

In talking with the data working group about a poster I was planning for the International Digital Data Curation Conference (IDCC) to be held in San Francisco in February of 2014, the working group members decided to create a poster about our data migration pilot project. Since the EcoPrairie project office and the campus library were willing to support their attendance at the conference as professional development, we embarked on discussions that made the data work involved in the pilot project explicit. We titled the poster 'Data Curation Issues in Transitioning a Field Science Collection of Research Data and Artifacts from a Local Repository to an Institutional Repository' (Kaplan et al., 2014a). The project collection and a vision of interoperability within a web of repositories were highlighted. Concerns with machine-to-machine connectivity as well as for human viewing and download, were shown to transform the EcoPrairie project data from an isolated collection to a publicly accessible resource.

The poster project required our working group to identify and articulate the goals, challenges, and big picture associated with data migration. While the information manager had a deep understanding of the requirements for the data to be submitted to the LTER data archive, the poster effort drew out and integrated our individual mental models, revealing a number of the differing embedded assumptions in the library collection and research project approaches.

Participation story 3. A data working group technical report

Since the poster project proved both engaging and productive, my suggestion that we next write a technical report about our data migration pilot project was met with interest. The group was not familiar with the format of technical reports so the purpose for such reports was discussed before beginning what became a year long, collaborative activity. The report titled

'Packaging, Transforming and Migrating Data from a Scientific Research Project to an Institutional Repository: The [EcoPrairie] Collection' (Kaplan et al., 2014b) was added to the online version of the project collection (EcoPrairie Collection, nd). The report prompted rigor in our definitions of terms and detailed consideration of workflows. It provided a place to document issues encountered and to formulate key lessons learned. In working on the written document together, we discovered points where we were misunderstanding existed and required not only discussion but also negotiation before moving on. The report effort brought clarity to our thinking as we addressed issues existing at the boundaries of the natural sciences, information sciences, and library sciences.

Participation story 4. Project closing data activities list

With a forward focus on building and maintaining a network, the LTER Network had not spent time developing decommissioning guidelines. Eventually data management concerns arose for EcoPrairie regarding the lack of guidance on data issues at project closing that impact a contemporary site and its personnel (see 4.1.7.5):

> Nobody said, hey this is what we expect of you from decommissioning except to say we expect your data be in PASTA … we had to figure out what closing entailed. And it wasn't just data management, it was also field cleanup. And there was also finishing up writing things and field studies that everybody used, sample archiving … and building a new capacity for future research. So those, we figured that out. (DM)

NSF considered the project's final two-year supplement as a road map for closure. Faced with addressing a number of data-related issues at closing, we decided to capture them in a list that eventually was referred to as a 'Project Data Migration and Preservation Checklist' that was included in the pilot project technical report (Kaplan et al., 2014b, section 6.3).

Disconnecting from a digitally connected network was not entirely straightforward; there were identities, data products, and emotions to consider. When a long-term member leaves a network, there are questions about update of community mailing lists and online content in light of site closing. Becoming a non-member after such an immersive experience makes leaving a community difficult. Just getting people to talk about the closing of a site incurs the same kind of sidestepping as in conversations about failure or death of a family member. Lines of communication are disrupted or become dysfunctional. During my study, there were many ways used to refer to the end of EcoPrairie funding and network membership such as sunsetting, closing, decommissioning, shut-down, discontinuing, funding withdrawal, and termination. In a history of ecology (Kingsland, 2005, p239), the first three LTER projects that closed are mentioned as sites that 'withdrew' leaving the mistaken impression that the projects chose to withdraw rather than having their funding withdrawn or terminated. Emotions tied to collaborative traditions, project loyalties, and data-related fears due to loss of the long-term project, were apparent when a project member reacted with 'watchdog' like intensity to a request for data from the terminated EcoPrairie project:

> In my mind I'm thinking, they want all the LTER data but they are not partnering or co-authoring or collaborating with the scientists … But then I realized the data are clearly in a public repository where they can search … and download the data themselves. And so

> I'm not a watchdog. I don't have to see myself as a watchdog because that is not part of my role. (DM)

With the accumulation of issues at closing, my participation as a sounding board for a list generated by the EcoPrairie information manager in the form of a 'Dear LTER' letter outlining ten issues or questions about site-related content within the network databases and websites. This letter was not sent officially but became the subject of brief discussion at the site. One issue related to whether the site information manager would be included in an upcoming information management meeting so as to be up-to-date on changing metadata requirements for data ingest. Concerns were emailed to the IMC chair which 'sent up this red flag alert' to everyone about site closing issues. Discussions between the community's Information Management Committee and the Network Office as well as between the Network Office and EcoPrairie project management occurred. It was decided these were issues to take before the LTER Science Council. The issues were recognized as a need for what is called 'critical procedures' in cases of site closing. Although the issues appeared as a brief item on the Executive Board agenda as "Managing data for decommissioned sites maintaining collaborations", later follow-up was not about the need that inspired the item but rather about dialogue with NSF aimed at defining site review criteria.

In hindsight, the Data Migration and Preservation Checklist in the data migration technical report (Kaplan et al., 2014a) could be expanded to include the subject of critical procedures as follows: 15. Review of network policy for discontinued member projects. This item would require at least four subheadings: a) continuing communication, b) meeting participation, c) presence in network websites, and d) capture of lessons learned in closing.

APPENDIX C. EcoRiver Timeline and Project Data Work

C.1 EcoRiver Timeline

1818 Illinois becomes a state
Emiquon area is a natural floodplain of a large river

Period 1
1858 Illinois Natural History Survey established (INHS)
1894 Biological Field Station established
Later renamed Stephen A. Forbes Biological Station (FBS, INHS)
1900 Reversal of Chicago from Lake Michigan (Chicago Drainage Canal)
1921 Levee built at Emiquon; Thompson Lake and Flagg Lakes at Emiquon drained
For agriculture and privatized hunting
1943 Illinois River major flood event (May1943)
1970 National Environmental Policy Act (NEPA)
Council on Environmental Quality (CEQ)
1972 Passage of the Clean Water Act by US Congress
1974 Water Resources Development Act by US Congress
Nine subsequent related acts 1976 to 2007

Period 2
1980-1986 Illinois Large River Long Term Ecological Research project

Period 3
1986 Upper Mississippi River Management Act of 1986
Twin mandates: economic development and river restoration
Upper Mississippi River Restoration Program established by USGS
Long-Term Resource Monitoring Program (LTRMP) established by USGS
1989 Illinois River Biological Station established (IRBS, INHS)
1993 Illinois River flood event (August 1993, Great Midwest Flood of 1993)

Period 4
1998 Emiquon Conservation Plan by The Nature Conservancy (TNC)
Worked with partner organizations to conserve biodiversity in the river
2000 Emiquon formed as Illinois floodplains restoration project
Nature Conservancy >7,000 acre purchase designated Emiquon
2001 TNC Emiquon Science Advisory Council established
2005 pumps stopped at Emiquon; rainwater (re)creates lakes behind levees
2005 Therkildsen Field Station at Emiquon (TFSE) planning begins
Nature Conservancy land; University of Illinois, Springfield building
Friends of Emiquon non-profit established at UIS
2007 First annual Emiquon Science Conference

Period 5
2008 Therkildsen Field Station at Emiquon established
2008 Prairie Research Institute (PRI) at UIUC created for transfer of four state surveys: Illinois Natural History Survey (INHS), Illinois State Geological Survey (ISGS), Illinois State Water Survey (ISWS) and Illinois Sustainable Technology Center (ISTC).
2010 Illinois State Archaeological Survey (ISAS) added to PRI

2010 NSF 2 year field station planning grant (M. Lemke PI, UIS)
2012 Science planning meeting with two LTER Data Managers invited (Nov 08)
2013 Education planning meeting (Feb 22)
2012 Emiquon "Wetland of International Importance" designation by Ramsar Convention

2013 TFSE initial ethnographic study of data work (March – June)
2013 March - Emiquon Annual Science Meeting; first Data Management Poster

2013 Illinois River major flood event (April 2013)
2013 Emiquon Rapid Grant funded by NSF (2013-2014)
2015 Illinois River major flood event (June 2015)
2016 Illinois River major flood event (April 2016)
2016 Emiquon Data Stewardship Workshop by TNC (March 21)
2016 IRBS Data Rescue Project with UIUC Archives (April 13)
2016 Emiquon levee gate completed for managed connection to the river (July 2016)

C.2 EcoRiver Project Data Work Expanded

This expanded description of EcoRiver project data work includes four sections: the role of data management, project collective data work, relations with partners, and stories of participation.

C.2.1 The role of data management

The role of data management was introduced at a field station planning meeting in 2012 where an EcoRiver tradition of partners working collaboratively on preserve management, monitoring, and research broadened to include data management activities. In considering a data management plan, a researcher explained the need for something more standardized than a spreadsheet approach to data management:

> To me a data management section means that it's a data library. So there's certainly data that you can use, and a reasonably good idea of some kind of quality control, that someone has looked through it. But then also it has the ability to be queried. You know ask things of it as opposed to a dataset of fifty-five different spreadsheets with twenty-five different formats and so forth. So there is some kind of uniformity to it. The fields match up, wherever the common element is, by GPS location, by date. And then that the units line up. …I know that's a lot to ask but lastly, and if possible, if the methods can't be standardized, at least they are stated someplace. So when I go back and I try to

> understand this guy's idea of organic matter estimation, and [compare to] the way I do it, maybe it's not exactly the same, but I have some idea of how to state what's going on … Standard methods of course are always better. (RS)

The idea of 'a data library' captures the notion of assembling data in the manner used for physical items. Data in a library conveys that it can be found when needed. Also, there is some expectation of quality expressed as 'someone has looked through it'. The need to identify common elements and catalog is understood though not expressed in the language of controlled vocabularies and metadata specifications associated with data work today. In addition, the mention of recording 'organic matter estimation' suggests the difficulties related to measuring a general biological concept in contrast to a physical parameter such as temperature. As a field researcher in the natural sciences, there is an understanding that field sampling often requires or is improved by adapting or inventing a method for the situation at hand. There is an appreciation that nature, the understanding of nature, and field methods will vary so the techniques used, locations, and conditions studied must be documented explicitly.

With a library, security and preservation of the data are implicit. Further, the library metaphor seems appropriate for scaling up from an individual to an effort involving visits over long time periods as well as collections of community, regional, and global extents. Whether with books on a personal library shelf or a public library shelf, easy access is expected. In the above response, however, there is no mention of data issues relating to the choices and packaging involved in moving data from the hands of a researcher to a data center. Instead, further discussion turns to the human element, that is, the need to coordinate with a changing set of individuals involved in longer field studies in order to maintain sampling, quality assurance, and quality control. With increasing ramifications of the digital era on research in terms of microchips in instrumentation, computers in the field, and storage devices at hand, it is not surprising that an EcoRiver participant recognized the need for new data practices. With increasing awareness of data sharing and access, new kinds of research questions are enabled in the form of grand challenges addressed by interdisciplinary groups One researcher's ability to conceptualize a new approach for working with data and some kind of title for a new data role tied directly to the library metaphor:

> You need some kind of curator or librarian. What do you call these people? (RS)

As work in the digital realm continues to expand, new roles and titles have emerged such as data specialist, data manager, information professional, and data curator (NRC, 2015; Mayernik, 2016). The use of 'curator or librarian' suggests an understanding of having services supported by a new kind of data worker. Perceptual readiness for a new category of work is evident though conceptual readiness is tied to existing positions in museums and libraries.

As spreadsheets were outgrown, the need to organize data on larger scales at EcoRiver led to next steps of development of databases and data systems. One EcoRiver researcher focused on the technical accomplishment of a centralized location with finished data products accessible online. In imagining a finished online product, the interim steps of developing new data work practices and new data roles were skipped over:

> Well I guess if I had to think about it [data management] conceptually, what I would think is that you have some type of a location. You know, URL or whatever that you go

> to and or maybe software that would combine and organize data from a lot of different studies. (RS)

Though data management became recognized by researchers at EcoRiver, there was also a reality check in terms of significant costs associated with data work, especially with so many field researchers dependent upon grant funding. One researcher noted that data management is "very necessary, but I don't know who pays for it". Indeed, the cost of collective data work today remains an unsettled issue.

Researchers, research assistants, technicians and graduate students carried out data handling at EcoRiver. The notion of a separate position for data management was absent. A research assistant at EcoRiver Field Station who collected and assembled heterogeneous, multi-variable data from several aquatic locations before creating a time-series dataset explained the work of a data manager:

> [As a data manager] I think you would be working with data that's more different. Everything that I'm working with is very, very similar. And I'm thinking if you are working with different projects, from different people, you are getting more input from different sources. (PS)

This is an individual who coordinated sample analysis across several departments and internationally as well. Increasing amounts and kinds of data were taken in stride although the work in managing the accumulation over time with constant attention to standardizing, error correction, and new products was not mentioned. Despite awareness of changes in the realm of data work, it is difficult for research participants used to handling the unknown of science and the contingencies of the field to imagine the need for new roles and new distributions of responsibilities for data work.

C.2.2 Project collective data work

Data management support: Initial planning

Data management for a project is easier to develop in theory than to put into practice with requires development of new language and new conventions. Prior to EcoRiver Field Station data management planning, one form of collective data work involved a set of subcontracted reports on key attributes and indicators (TNC, 2006). With data management entering the discussion, an initial online investigation of existing materials and plans resulted in an ambitious, optimistic EcoRiver field station planning grant statement that embraced the need for data management but lacked practical detail and timeframes:

> A major part of the proposed planning will be establishing among the partners a common protocol for collecting, processing, analyzing, and documenting data products; and for archiving, curating, and publishing these products. (RS)

In this statement, there are embedded assumptions about protocols as readily enacted, data products as easily identified, and archiving as an established process. The community intent to engage in collective data management is evident in the plans. Indeed, the plan is taken as an opportunity to enhance local capabilities by engaging with contemporary digital issues. Despite

differing concepts of data management, EcoRiver Field Station and EcoRiver partners had begun envisioning data management options and scenarios. Two researchers agreed on a five-year, weekly time-series as a dataset:

> The five-year dataset that [they] have been keeping on the lake, no doubt. I mean that absolutely, to me, that is the data gem that has been created at that field station. (RS)

> I think that five-year dataset would be a dataset … it would be all the DNA, and all the physical, and all the chemistry, and all the plankton. That would be the dataset. (PS)

These researchers and others subsequently referred to the data as 'the physical and nutrient data' and 'the physical dataset'. Confusion arose because the 'physical dataset' category included both temperature and biomass where biomass is a biological rather than a physical or nutrient measure. Indeed, participants described the five-year dataset at various times as having two, three and four data streams (Baker, 2016, Figure 4.1). Development of a workflow diagram revealed the dataset as an aggregation of measurements, sometimes referred to individually as datasets, but bundled into a single spreadsheet as a final data product.

Oversimplified language hides the design aspects of data work. An imagined next step with this data was described as 'into the database'. Putting data into a database introduces a variety of structural arrangements using tables and relations. The phrase seemed to connote the notion that existing data and information issues would be solved by the move of data into a database. To open up this database view, an information system view was presented side-by-side with a database view at one of the EcoRiver annual science meetings (Appendix C.2.4, Story 3). Three major elements of data systems were portrayed: a) data gathering and format transformations required for collective data assembly, b) backend database functions that handle dataset metadata, project content such as personnel and publications, as well as standardizing content such as codes and methods, and c) front end interfaces that provide web access and download of data. Though these elements are fundamental to understanding and planning collective data work, they were found in practice to be outside the general knowledge - and interest - of most research scientists in the natural sciences.

By the end of the EcoRiver Field Station two-year planning grant, general interest was accorded data management. Leaders were able to speak to the benefits of supporting data management, of 'having one person… well, a data manager'. There was both a sense of need and willingness about addressing new data requirements. This situation was influenced by a number of factors: the interdisciplinary approach on a small college campus with expertise in ecology, computer science, and media studies; geographic proximity to a research study location managed by a non-profit pursuing restoration guided by scientific management; a tradition of taking in stride new factors in an ecological landscape; the relatively small scale of a field station with a minimum of legacy technology issues; and the serendipitous presence of one researcher having past experience with collective data management at LTER sites together with other participants having experience with a distributed network. In addition, there was a purposeful inclusion at this teaching university of outside advisors including experienced information managers.

In the absence of a data manager, the research data plan led to a pilot project from which partners gained collective insight. In the two-year multi-partner research grant, attention was given to data management as a collective process to be addressed with a new approach. First, the work of assembling data was recognized as time consuming with different participants having

different timeframes for engagement resulting in difficulties of coordination and communication. Second, the assembly of data was recognized as requiring three staging arenas for the different versions of the data: a) original: the data submitted by researchers with upload permissions; b) original copy: a preserved copy of the original data; c) master: a final version of the data that has been checked and cleaned by the data manager. Finally, a decision to ingest the master copy of data into File Maker Pro software that was familiar to several local participants had a number of ramifications. It provided an opportunity to comment on three differing approaches to data organization and description: file systems, relational databases with a single key, and relational databases with multiple keys. Ultimately, the organizing and disseminating of data on a standalone platform was recognized as inadequate. Toward the project end, EcoRiver participants worked with the university central information technology department to install the database on a server in order to establish project-wide access to data rather than working with distributed copies of a database. This step revealed the need for expertise in database programming to continue with this approach. In this pilot project, a number of partners demonstrated readiness to assemble and share. Major hurdles were deciding how to structure data products and what mechanisms to use for collective data management when needed infrastructure, support personnel, and funding were not in place to support such data efforts.

One EcoRiver participant in 2012 envisioned an overview log of activities, an assembly of information about projects. The concept of such a data catalog arose again at a community-wide workshop in 2016 (Appendix C.2.4, Story 2) attended by the individual working within EcoRiverOrg that carried out management of permits, for applicants requesting permission to sample at EcoRiver. A requestor filled out paper forms and the information was transcribed into an Excel spreadsheet. The topic arose when it was realized that the data collected by the owner/managers of the preserve would not be in the catalog since EcoRiverOrg projects were not included in their permitting process. The project catalog emerged conceptually as a way to tie data products back to their origin with the sampling permit number serving as a unique identifier.

Local systems support: Individual tasks

The new field station seen as a sampling staging area, a club house, and an education venue. Though plans were to provide infrastructure for the technical, social, and outreach aspects of fieldwork, data arrangements were not addressed. From a field data collecting perspective, the EcoRiver Field Station represented a new organizational entity supporting fieldwork at EcoRiver:

> What [EcoRiver Field Station] offers is the field equipment on site … we have it so that the prep can be done here and the post sampling processing. So by the time they get in every week, things are stable. They [the samples] are either on a filter, they are frozen, they are preserved, bam, done. As opposed to when we used to sample a big part of the data [by] hauling our boats out, sample, and then people are very exhausted. We'd get back [to the university] and then hand all these samples over to one group of people who are doing filtering, one group of people who are doing the processing. And it was pretty exhausting. (RS)

In discussing infrastructure for field data, one EcoRiver researcher mentions some of its dimensions – technical (machines), practices (protocols), and social (connections) dimensions:

> Well I would say there'd be machines, protocols, and a person, at least one person, designated at the field station. Then there would be machines, protocols and people here at the university. And then there would be a virtual infrastructure, so that people from outside can access the data that is essentially received and processed at the field station and archived at the university, and perhaps somewhere else in the cloud. So I see it as a three staged, three-piece model: the field station, the university, and the Internet virtual presence. (RS)

Though a multi-component framework is described for data movement from the field to the university to the Internet with web delivery to the pubic, specifics about gathering, managing, organizing, documenting, and packaging were not mentioned.

EcoRiver technical support was available from a number of organizational entities. EcoRiver Field Station had ties to the campus-wide IT unit that provided services on a limited scale such as hosting servers and shared storage. As a teaching university, there was not much call for preserving scientific data though university librarians were aware of recent work on data curation at other libraries. Institutional repository services were provided by a sister campus library but were little used by EcoRiver researchers. EcoRiver had ties to ecological data work occurring at the Prairie Research Institute at UIUC. The institute supported a data portal for an array of databases holding field data. The databases were managed independently within separate laboratories. Finally, the UIUC central library was in the midst of launching a data repository effort that held potential for addressing some of the data needs of the UIUC Illinois Natural History Survey field stations within the Prairie Institute (INHS, nd).

Individual and collective planning

A variety of experiences with data assembly involving various data collections and data repositories existed across the EcoRiver community of researchers. In speaking of a nearby laboratory of another researcher, one participant mentioned familiarity with the neighbor's research but a lack of familiarity with their approach to data organization:

> I know they generate a lot of data, but I don't know how they manage it. (RS)

This was not an unusual for EcoRiver researchers facing increasing amounts of data together with higher expectations and more collaborative projects. In exchanging data within a laboratory or with a colleague, one researcher established laboratory metadata practices to help guide interpretation of the lab's data:

> Most of our stuff is actually just in the Excel files … so that's what I would send you. Before I send it to you though I would probably add another sheet in the front to try to put as much orientation on there so that you could interpret it. I think I have the information I need, but you know, obviously you have to look at it from the perspective of somebody who didn't do the sampling and doesn't have that recollection. (RS)

One participant associated with the EcoRiver Field Station considered even a small digital collection as a data repository.

> I guess this [a set of spreadsheets] is a data repository. And when I need to connect up

with other people, Dropbox is a data repository. (PS)

Another EcoRiver participant described the importance of planning arrangements for data assembly that ensure security and public access when explaining what a data repository is:

> a place where you can store, archive, backup, a secure location for data. When I use the word repository I think of it as multiple people and institutions, multiple people have added their own data to it. … You could be a repository and archive without being publicly accessible. But this one I would want to have some provision for making it public. (RS)

A researcher identified the data repository as an 'essential place' adding the proviso that data be available online with access permissions appropriate for larger, distributed audiences:

> a location, may be online … basically you can store the data in this essential place. And scientists can access, with certain permissions, access the data. (RS)

A participant working with computer systems at EcoRiver Field Station showed awareness of some of the many issues that accompany an assembly of data when defining a 'data repository' as having multiple stages of repository development (data description, archive, access) and multiple stages of data (raw, cleaned) as well as considering the idea of a 'depository' as a field station service:

> I think we would want the cleaned up data rather than this preliminary data. But their data is submitted to our archive with either a default or a negotiated date at which we can make it public. And that public [access] is through the field station's website. And it's a big data depository. At first it will simply be archived and you have to know the day and the person. Some day we will make it searchable for people. I would really like to have every dataset that is in that archive have as part of its metadata the physical locations by GPS coordinates, latitude and longitude coordinates. (RS)

These comments by EcoRiver participants show individuals involved in collaborative research envisioning a data repository for assembling data prior to having in-depth experience with collective data management, data systems, or archive partners.

Website for a variety of partners

The EcoRiver website supported by a museum partner, indicated awareness of presentation of material online as an integrative mechanism that contributes to a project's narrative (EcoRiver Outreach, nd). The website underscored the mission of conservation, reported on engagement with volunteers, and provided a list of partners and affiliates. Navigation tabs led to sections on history, research, news, events, recreation, and resources as well as volunteer and visitor services. The research page, as one element of the larger story of the preserve, contained high-level announcements about activities, grants, and publications. The partner list provided links to partner websites including the new EcoRiver Field Station.

Early enthusiasm with the EcoRiver field station web presence gave way to the realities of maintaining content as well as real-time connections with instruments in the field. The website

was hosted on campus servers. Three on-campus individuals in different departments managed it. Their work with the new field station website arose from an interest in social media and effective communication with the public. EcoRiver Field Station participants spoke early on about the website as hard to keep up. The website provided basic information about the field station as well as research projects, publications, and conferences. A data tab initially provided output from a direct link to an instrument in the field that was their first experience with data streaming. After redesign of the website, the data page stated: "This page will host links to current and historical datasets collected in and around the EcoRiver preserve. Have data to share? Send us a note. Check back for updates." The data tab became a placeholder for data products to come.

C.2.3 Relations with partners: Multi-scale partnership

EcoRiver participants coordinated data informally within labs as well as in conjunction with funded research projects and EcoRiverOrg preserve-wide efforts. The partnerships illustrate the interweaving of science and data efforts required to support decision-making about use of natural habitats.

A data stewardship workshop sponsored by EcoRiverOrg's statewide science office in collaboration with their regional office occurred in response to growing awareness of data needs at EcoRiver (Walk et al., 2016). Three data work categories associated with collective data management efforts were identified during workshop discussion:

- System administration (hardware, security, and application support)
- Programming (application and database support, web design)
- Data management (data assembly, metadata, web content)

Though often existing in practice as three separate roles that represent what was called a 'Minimum Data Team', available funding and expertise could lead in practice to one or more individuals performing this work depending on the skills of available staff. Workshop co-leaders started by identifying a data team with a system administrator in Chicago, a programmer specialized in GIS in Peoria, and a proposed information management position also to be located in Peoria at the state science office (Walk et al., 2016; Appendix G.3). With such a team in place, participants considered the EcoRiver as a regional project that would serve as a prototype for other regional efforts in the state.

Workshop discussions of data work and data teams frequently ranged across data needs of the statewide organization, of project-specific regions, and of individual participants. The plan outlined for a statewide data team was envisioned as raising the level of data management capacity for not only all EcoRiver regional projects but for partner organizations as well. This hierarchical structure of state headquarters, regional projects, and member organizations suggested a need for further work on data policy and governance issues. Specific roles for EcoRiver community members, who are an important source of data for the regional and state offices, were not discussed. These community members from research units in their own primary organizations, were already involved in an assortment of legacy and ongoing data arrangements, projects, and practices.

In addition to collaborating at the regional level, various EcoRiver partners worked in a myriad of partnerships at other scales such as large-scale water basins and the globe. Government agency partnerships enabled much of this work at larger scales. For instance, beginning in 1986, funding and overall responsibility for the Environmental Management

Program was assigned to the U.S. Army Corps of Engineers (Corps), which transferred funding to the U.S. Geological Survey (USGS) to administer and implement a federal-state partnership program called the Long Term Resource Monitoring Program (LTRMP). LTRMP grew to include field data collection, analysis, and applied research. A number of early science and information center efforts combined to become the Upper Midwest Environmental Sciences Center (UMESC) that provided science and data services for the LTRMP. Work is carried out in cooperation with natural resource agencies from the six states along the Upper Mississippi River: Illinois, Iowa, Michigan, Minnesota, Missouri and Wisconsin. Each state supports a state-owned and operated field station including one of the EcoRiver partners at Havana, an INHS field station supported by federal funds with supervision by the US Fish and Wildlife Service and the Council on Environmental Quality.

LTRMP, with its multi-agency, multi-state collaborative arrangements, was only one network with which EcoRiver partners interacted. An international Large Rivers Consortium (Sparks, 1995) was in a position to contribute data and information in terms of scientific assessment of river well being. Further, the National Great Rivers Research and Education Center (NGRREC, nd) was formed in 2002 to study the ecology of big rivers and their environments through a partnership of the Illinois Natural History Survey, UIUC, and Lewis and Clark Community College. A bridge to policy was evident when critical information was provided during a recent clash between economic interests of commerce and environmental concerns about detrimental impacts of river navigation on habitat health (Sparks, 2016).

C.2.4 EcoRiver stories of participation

The following four stories of participation and partnering cover a period of five years (2012 to 2016). They summarize some of the actions relating to the introduction of data management to the community of partners during this study by: 1) inviting information managers to planning meetings; 2) opening the door to data management discussion and study; 3) including data specialists at community science meetings; and 4) conducting a data stewardship workshop for project partners.

Participation story 1. EcoRiver field station planning grant meeting in 2012

An NSF field station planning grant for the new EcoRiver Field Station included a section on data management. The planning grant supported a science meeting in November 2012 for research participants that included attendance of two information managers, one currently an LTER information manager and myself, a previous LTER information manager. The invitations were an example of the diffusion of LTER experience. A researcher affiliated with EcoRiver Field Station had received his doctorate while working at another LTER site so was exposed to the LTER culture of embedded data management. During workshop planning, he suggested that an LTER information manager be invited to the meeting.

At the workshop wrap up session, each participant including the two data managers shared suggestions about top priorities for the EcoRiver Field Station. The data representatives coordinated their two suggestions: 1) recognize data management by designating a data manager and 2) begin collective data work with the EcoRiver sampling permit process not to control but to document. As a result, there was discussion of station-related data management issues and options for a part-time role of data management filled perhaps by a station manager or a field technician. A field technician was recognized as having the advantage of familiarity with data

entry while a site manager was considered as having a position with more stability than techs "who will kind of cycle through as they work on their degrees".

Participation story 2. EcoRiver Field Station data management study report

An ethnographic study of EcoRiver Field Station was carried out during the summer of 2013. Drawing on interviews and participant observation, a class paper for a UIUC qualitative research methods class was written that provided an overview of the state of data management at the station. The material was rewritten subsequently as a report with revised text and an expanded timeline. Participant quotes previously gathered into appendices were incorporated into the report itself. In addition to providing a brief history of the site and its science as part of a large river system, themes of data management, data sharing, and infrastructure were discussed. In the years between 2013 and 2016, my continuing collaboration with EcoPrairie provided opportunities for member checking including having several EcoRiver participants provide feed back on the full report (Baker, 2016).

Participation story 3. EcoRiver annual science conference presentations

Over the years at EcoRiver, an annual science conference provided a forum that brought together the diverse research activities at the preserve and provided an opportunity to work with participants to present information about data management. I partnered with a computer scientist and a research assistant associated with the field station for a poster in 2013 to present two contrasting views of a data system: a technology-oriented database view and a science-oriented information system with six features called out (Baker et al., 2013). In 2014, in addition to presenting a poster titled 'Planning for Data Management at EcoRiver' (Baker et al., 2014a), an invited talk was given on 'Developing a Data Management Strategy: Infrastructure for the EcoRiver Partnership' (Baker, 2014b). The poster and talk introduced conceptual models for data work and data infrastructure development. A 'continuing design' approach was presented in a poster for the annual science conference in 2015 in response to a decision made by one group to put data into a non-relational database application. Since query in addition to access was likely a feature that would be needed in managing EcoRiver's heterogeneous data, planning for data export was discussed so data could be imported into a relational database in the future when additional data expertise was available.

Partnering with an information science student and an EcoRiver research assistant in 2015, a poster titled 'Data Management: File Systems, Databases, and Metadata' reported on differences between file systems, databases, and datasets with metadata (Baker et al., 2015b). Among the participants who stopped at the poster, lack of exposure to concepts needed for data planning and collective decision-making was evident. The poster introduced vocabulary pertinent to data management and ties to digital infrastructure that is integral to today's scientific enterprise.

Participation story 4. EcoRiver data stewardship workshop report

An EcoRiver Data Stewardship Workshop in 2016, co-led by two researchers affiliated with EcoRiver Center and myself, was summarized in a report (Walk et al., 2016). Surveys conducted before, during, and after the workshop gathered views on data work. Findings included responses on the importance of data access and integration (high), high priority activities (permit catalog and dataset catalog) and an evaluation of the workshop ('informative' and 'overwhelming'). Rather than calling it a data management or data curation workshop, the

title ‘data stewardship’ was chosen to parallel the EcoRiverOrg management board’s familiarity with the concept of land stewardship. By analogy, the aim was to convey the responsibilities and complexities associated with data stewardship. Two significant data activities were identified: creation of a more comprehensive dataset catalog and discussion of a complete list of sampling locations, both important to project transparency which in turn enhances collective data work. In addition, an information management position, planned beginning in 2014 at the annual science conference, was included in the organization’s five-year plan. An information management job description was shared with participants at the 2016 Data Stewardship Workshop (Appendix G.3).

Workshop preparations created a unique opportunity to carry out translational data management. Pre-workshop preparation included biweekly virtual meetings of one to two hours for a period of six months. Frequently, I prepared a series of 5-10 slides containing information about data management. We discussed and reformulated slides while identifying topics, outlining an agenda, and clarifying materials. My co-organizers used this information to develop community appropriate presentation materials for the workshop. From these sessions emerged strategies and topics salient to EcoRiver participants. The final workshop report represents an extreme form of traditional ethnographic ‘member checking’ in that one of the community co-leaders produced the first draft of the workshop report, and then co-leads joined in writing, reviewing, and editing the report. Notably, we each assimilated the material into our personal views of data work so were all comfortable presenting the data management materials at the workshop. One participant’s final comment indicated the workshop conveyed the multiple dimensions of data management: “I came with a preconceived notion, but it’s more complex than I realized. I didn’t know about it all.”

APPENDIX D: AtmChem Timeline and Project Data Work

D.1 AtmCenter and AtmChem Timeline

Period 1: Early NCAR Planning and Operation
1957 International Geophysical Year (1957-1958)
1959 Blue Book on 'Preliminary Plans for a National Institute for Atmospheric Research'

Period 2 Atmospheric Chemistry identified as a key program
1966 ACD, Atmospheric Chemistry Department organized
1967 -1969 ACM, Atmospheric Chemistry and Microphysics
1967 NCAR brought together instruments and people
early GATE – talked about distributed archives
1970-1973 ACD, Atmospheric Chemistry Department
1974-1979 AAP atmospheric analysis and predictions division
AQD atmospheric quality division
1975 Computing facility for field operations
1980s NASA Mission to Planet Earth (MTPE) renamed first NASA Earth Science Enterprise (ESE) and then NASA Earth Science
1980-1984 ACAD, atmospheric chemistry and aeronomy Division
1985 ACD Atmospheric Chemistry Divison
1994 ATD atmospheric Technology Division annual reports (ends in 2004; see EOL 2005)
1995 NCAR Information Infrastructure Technology Application (IITA), organizational unit
- Scope: Enterprise IT organizational infrastructure
- Outcome: IT strategic plan for organization

1995 Field catalog, first digital version
2001 NCAR Data Management Working Group, cross-cutting group
- Scope: Scientific computing interoperability of existing capabilities
- Outcome: Community Data Portal

2003 Field catalog, last print tech report version

Period 3: EOL established and expanded mission
2004 Reorganization to create EOL approved (Apr)
- Merge of FODM with ATD transition planned (June 2004 – June 2005)
NCAR/ATD Atmospheric Technology Division
UCAR/JOSS/FODM Joint Office for Science Support (JOSS)
Field Operations and Data Management group (FODM)

2005 ESSL, Earth and Sun Systems Laboratory, formed
with divisions: ACD, GCD, HAO, MMM, TIIMES
2005 EOL with Computing, Data and Software Facility (CDS) formed (June)
- EOL expansion into biogeochemical monitoring (from JOSS)
- EOL Mission Three D's: Development, M, and Data Services

2007 EOL aims to further develop EMDAC – EOL Metadata Database and Cyberinfrastructure inclusive of CODIAC, web interface to data holdings
2008 EOL strategic goal 4: Provide robust, accessible, and innovative information services

and tools
2009 NESL, NCAR Earth Systems Laboratory, formed
with divisions ACD, CGD, and MMM Divisions
2009 EOL Imperative IV: Provide comprehensive data services, open access and
long-term stewardship of data
2010 Field catalog next generation (overlays)
2010 NCAR Committee on Data Citation (CDC), an adhoc group
- Scope: Scientific computing research digital services
- Outcome: NCAR-wide DOI Services coordinated and documented

2010 EOL Mission Expands
- Four D's: Development, Deployment, Data Services, and Discovery

Period 4: Further organizing for collaborative science
2011 NESL NSF Site Visit Team review (May)
2012 Community workshop reviews NESL including ACD; propose ACCORD (Feb)
2012 ACCORD Data Management and Integration working group formed
2013 ACD and NSSL annual reports do not mention ACCORD
2013 Ethnographic Case Study (2013-2014)
2014 NCAR Data Stewardship and Engineering Team, formally recognized NCAR-wide group
- Scope: Scientific data stewardship
- Outcome aim: Common discovery of resources

2014 ACD Website updates of data catalog
2014 ACD and NESL Strategic Plans with section on ACCORD
2015 ACD becomes ACOM (Atmospheric Chemistry Observations and Modeling)
2015 NESL dissolved with ACD, CGD and MMM becoming laboratories
2016 National Academy Press review of atmospheric chemistry (NAP, 2016)

D.2 AtmChem Project Data Work Expanded

This expanded description of AtmChem project data work includes four sections: the role of data management, project collective data work, relations with partners, and stories of participation.

D.2.1 The role of data management

AtmChem is a research unit where researchers had established practices for managing their own data in the field and the lab prior to submitting data to existing archives. They had familiarity with computer applications and processing of data files. Data work at AtmChem included a suite of overlapping tasks and skills associated with data use and data archive, tasks typically performed by participants with differing backgrounds in science, engineering, and data management (Mayernik et al., 2014).

A data management role as distinct from that of a scientific data generator, systems administrator, and systems or software engineer did not exist within AtmChem. Research scientists' focused on the complexities of atmospheric chemistry had little interest, training, or time for considering collective data management and data preservation. In seeking to explain the technical staff's difficulties in communicating with researchers, a software engineer observed

that at research divisions "they don't even have dedicated data managers there". As a result, in dealing with data issues within this organizational research unit, the AtmDM group interacted with a changing cast of AtmChem participants resulting in little continuity for data discussions.

Though titles such as 'data manager', 'information scientist' and 'information professional' were not in use at AtmCenter, a data management role was created within AtmCenter's library. With a growing awareness of digital scholarship and repositories, an information scientist was hired with a title of 'Project Scientist and Research Data Services Specialist'. The title was designed to provide recognition of the leadership envisioned for the position and to convey status comparable to that of software engineers and managers. One technical staff member in a research unit described the library move not as a territorial encroachment but as supportive of institutional data efforts:

> In putting on my bigger picture hat and looking at different groups across [AtmCenter], I look at the library as being one mechanism that might allow cross-institutional coordination with data. (TS)

Previous titles for the heads of data and technology groups included 'Facility Manager', 'Manager of Data Support Section', and 'Associate Scientist'. The new specialist position bridged a research-technical divide populated by field scientists and research scientists on one hand, and engineers and computer professionals on the other hand. As one systems professional noted:

> We're working with pieces of AtmChem, and it's the same thing in other parts of the institution. … Maybe I'll be surprised, but I suspect you'll find it's a bunch of these little islands … I know they don't have things like vocabulary and coordinated efforts in data management. (UM)

In the staff directory, the library position was identified as associated with interests in "metadata practices and standards, data curation education, data citation and identity, and social and institutional aspects of research data". This hire by the library has contributed to organization-wide data work despite the challenges of establishing data contacts with each organizational unit.

The role played by data intermediaries was made visible by an Institute of Museum and Library Services (IMLS) grant co-led by library researchers. The project was designed to provide experience through internships that brought students from the Information Sciences to AtmCenter. This exposed researchers and engineers to a variety of contemporary approaches to data work while, at the same time, exposing students to the complexities of managing heterogeneous data, the realities of fieldwork, and the issues associated with observational data use in modeling. Posters and publications illustrate this project's findings (Mayernik et al., 2015; Thompson, 2015; Thompson et al., 2015; Baker et al., 2015c; Palmer et al., 2014; Palmer et al., 2013a).

D.2.2 Project collective data work

Project collective work is described in the following sections on shared platforms, data storage data assembly, and websites.

Local data management: Sharing platforms

Data collected on flights carried out during an AtmChem field campaign was saved on shared aircraft platform systems or on individual computers and storage devices. When AtmDM representatives were in the field supporting a project, they provided storage and networking for sharing and aggregating data in a central location. Both in the field and in the local arena, the timing for data releases for project and public access to data files was recognized as a critical factor not only for data sharing but also for quality control. Within projects, a standardized three-category nomenclature was used by these participants tied by use of a common platform in the field: field, preliminary, and final versions. AtmDM in participating in fieldwork associated with aircraft and instrumentation, developed methods for handling multiple versions of data. Early versions of data were made available to project participants. After undergoing post field processing and analysis within the context of other project data, a final version at the archive would be open to 'the public'. Timing was coordinated carefully:

> The field data we put out 24 hours later [after the flight], that's useful. During the project and sometimes in the field, they [the project leads] decide the date for preliminary data and for final data. There's some deadline that we have to meet. That preliminary data varies a little bit. It's usually around six months. (RS)

And data sharing within the project provided opportunities for data checking:

> So we provide the data in the field, where you've taken it off the plane and done a quick calculation before providing the data. Then there's some deadline to provide preliminary data; we provide that out to the group. That's often a significant step up from the field data ... you get a chance to look at the data, find bad data within the set and remove it and do at least some early analysis to improve that data. And the preliminary data is good enough for people [in the project] to take that data and start to do some more advanced analysis knowing that they may need to replace some of these datasets. In fact, if they are going to publish anything, they do need to replace the data with the final data. So preliminary data is within the project because you don't want to trust people outside the project to understand that. … And then the final data ideally is really final data, as good as you can provide from the measurement and that's what is released to the public. (RS)

Since field campaigns consist of multiple flights, perhaps one flight a day for five days, researchers spend time in the field determining whether their particular instruments are in working order or need repairs before the next flight.

> … We use the aircraft data to go back and model our data and compare our data to the model for a variety of reasons, the main one being when we're in the field to make sure we're not totally off base by an order of magnitude or something. … And so sometimes trying to understand the air that we're flying through, whether we're in the stratosphere or the troposphere or whether we're in a pollution plume. Aerosol data, that's another one I look at. And certainly other radiation measurements, the radiative transfer, to understand what is happening. (RS)

Data from the instruments of other project members is used for understanding field conditions and to assess how their measurements fit within this context. Early exchanges of project-related data and model results among participants are also helpful to individual investigators in identifying instrument malfunctions, data processing glitches, and possible outliers in their own data.

Local systems support: data storage and exchange

Local file storage and exchange was designed to meet researcher needs for ease of data access in targeted, loosely managed situations. Within their division, AtmChem supported a technical group staffed by three systems administrators having an in-depth familiarity with technical needs across their unit. These technical specialists oversaw the digital systems and applications that undergird the division's technical data infrastructure. They provided support for backups as well as a variety of file storage and exchange mechanisms important to the highly collaborative work of the unit's members with others within AtmCenterOrg and within other organizations. These technical specialists develop and maintain a variety of approaches to data access that vary depending on factors such as file size (e.g. use of high-performance storage system (HPSS) tape), internal data sharing (e.g. use of shared server storage), fast transfer to external destinations (e.g. use of file transfer protocol (FTP), and contextualized links with low overhead (e.g. use of the web). Support for a local systems group within the division reflected the priority given to exchanging data files and to preparing the 'final' stage of data for preservation and public dissemination.

An AtmChem technical specialist emphasized, however, that the storage space was considered a temporary 'stopping point' rather than a 'final destination'. Systems administrators were clear that the storage space made available within the division was a place where data could be put until it is moved somewhere else:

> We sort of deal with things situationally. So we do have an FTP server. I let people put files there until it fills up and then I nag them. Or until the disk fails. But the understanding is that is a stopping point not a permanent point for the data. We try to make sure that is understood. (TS)

Researchers, however, did not necessarily understand the notion of a 'temporary stopping place'. Stories of a disk crashing or file forgotten after a period of non-use underscore the difficulties either distinguishing temporary storage from permanent storage and of making permanent arrangements for all field campaign circumstances.

An example of local technical support as a complement to larger-scale services and facilities was observed in a response to an immediate data need. When a delay occurred in deploying a cloud storage system organization-wide, AtmChem technical specialists continued to receive inquiries about options available for distributed data access. After investigating temporary options for the division, the specialists identified OwnCloud as a web-based, open-source, self-hosted, and user-friendly product. It was deployed for the division and informally available to the whole organization. It was seen as an interim alternative during the period of waiting for a larger-scale 'store and forward' service under investigation at AtmCenter. A local systems manager contrasts their immediate response to researcher needs with an organization-wide response:

> I'd like to think it's a strength that we're able to spawn off efforts at the time that they are needed, relatively quickly, because we don't have really dictatorial management schemes in place. So the scientists are able to really drive the decision-making and to me that's kind of what ought to happen. It's sometimes at the expense of larger organization, at the expense of being able to organize things efficiently. (TS)

The cloud solution was of immediate use to a group that wanted to gather data from computers of research colleagues in another country where few technical resources were available. Student participants initially established remote login to a desktop, next shifted to the simpler solution of Dropbox until data overflowed available space, and then used an organizationally licensed Box solution until data limits were exceeded once again. Finally, they turned to OwnCloud. The OwnCloud option was seen as part of a process:

> When you develop something at the grassroots, like we've developed OwnCloud, and its successful enough, we can sort of take a next step and that's to get institutional support for it. So that would mean [a bigger IT department] providing it on their virtual machine cluster. (TS)

This example of a local technical support group's action, undertaken as a temporary measure, represents an experimental prototyping that can inform decisions about deployment at larger scales. It illustrates the rapid responses possible within organizational units and shows an adaptive mindset comfortable with interim designs for use until an alternative capability is available. In some cases, these rapid responses also contribute to identification of data needs that may not be recognized yet in other units.

A senior systems manager explained that a number of systems groups at AtmCenter operate separately but coordinate in an informal manner similar to a user group in contrast to advisory committees. The systems groups provided a critical mass of technical expertise and capacity within a number of the organizational units. Systems managers at AtmCenter were a long established, loosely knit community that swapped stories as part of an informal but effective communication network across research units in a manner Orr (1996) described as central to the workaday activities of those who administer and maintain technology. The success of technology arrangements within the organization rests on the trust established with and by such technical specialists. Within an organization, these groups are agents providing situated responses to individual as well as project data needs. One system administrator explained their 'working relationship' with three kinds of infrastructure:

> There's this kind of culture and it began very informal, very grassroots, none of it really directed by management. It just sort of evolved. But sysadmins tend to work with each other because mostly what we do in the day is solve problems. And so if a problem is difficult to solve, then we're going to pull in all the resources that we can, including our fellow sysadmins. And so that fosters a kind of natural working relationship … and then the administrative infrastructure and the computing infrastructure and to some degree the web infrastructure are really used to working with each other among the divisions. (TS)

It was observed that many project participants lumped together as 'technical' what is here distinguished as 'administrative infrastructure', 'computing infrastructure', and 'web infrastructure'.

Multi-project data archives

When AtmDM was the designated project archive, a software engineer was designated as the project data contact to assist in submitting data to the system. When NASA was the designated archive, there was an automated upload interface as well as a research scientist who served as data contact.

> There's a website with a formal submission process. NASA generally has that. This last project that we were on [with AtmDM] you had to go to the room with the local network and upload it to a drive. There were some data managers there so you could also give them a stick and say here's the data, put it into the set.

The NASA upload process included a format check of data files that were required to conform to the **I**nternational **C**onsortium for **A**tmospheric **R**esearch on **T**ransport and **T**ransformation (ICARTT) file format specification (Aknan et al., 2013; footnote: guidelines). The specification is named after the ICARTT campaign in 2004 where it was adopted to facilitate use of shared data. One AtmChem researcher explained about formats:

> [It varies] project to project. For some projects, it's required … the ICARTT format. For us [in our instrument group] it is, because our processing routine has that [ICARTT format] at the end anyway. We also spit data out for our tools in a different format. (RS)

As a well-established format in the atmospheric chemistry community, ICARTT takes multiple forms depending upon the kind of data, e.g. tabular, spectral, or images. An example of an AtmChem project statement of data sharing that appeared in a pre-field plan states: "To allow consistency with previous measurement campaigns, the ICARTT file format (modified NASA Ames) will be used. The ICARTT format is a text (ascii) file format that is easily produced and used by most investigators." The requirement that it be 'easily produced and used' precludes using another well-known array-oriented standard for satellite and modeling data consisting of a climate forecast set of software libraries with a self-describing format called Network Common Data Form (NetCDF). One AtmChem researcher explained that atmospheric chemists find NetCDF with 'the software and binary' to be 'overkill' for what they are doing. Some researchers limit their worked with NetCDF to extraction of the variables from shared data and then worked with the extracted data using local formats and applications.

Websites for many projects

The AtmChem organizational unit website was overseen by a web specialist who was a member of the unit's systems group. With a background in atmospheric modeling, this web master oversaw the website design and content for individual projects as well as the unit. Web work relating to project data was carried out largely in response to requests from researchers. The representation of projects and data underwent several changes during the time of this study.

Two kinds of project web pages were created in conjunction with a field campaign: first, a science-driven web page that evolved over the research timeframe and second, a standardized,

archive-driven web page. Large atmospheric research projects with participants geographically distributed often used a research project website as a mechanism for communication about logistics and document sharing. The formats of the project websites varied depending on the project principal investigator, the project participant hosting the website, and collective use of the website. AtmDM, as a self-described 'archive for field projects' with experience providing data support services for field projects, worked to capture science-driven project web pages together with other project-related web pages and materials. In order to browse easily across many projects and to organize project materials, AtmDM maintained a post-project website standardized via a template used within a content management system. The standardized template included a brief project description and scientific objectives appeared under a project logo and photo. Accompanying categories of information included links to data access, data documentation, publications, documents, meetings, facilities, logistics, and education/outreach.

Issues arose with the two kinds of web presentations due to the sensitivity of scientists to maintaining project authority. The post-project website, referred to as a 'final project website', contained a section called 'Related Links' that provided links to the 'science project website' as well as links to websites of related field activities including education/outreach sites and other sites that may be hosted outside AtmCenter. The final project website effort was considered by AtmDM as adding value to data interpretation by providing context often lost when project web pages were hosted elsewhere. Constructing a standardized project website and capturing project-related content proved of growing concern to AtmDM due to its demands on time. Web support was sometimes requested as part of a research project's planning but other times, in the interest of completeness, a template was created after-the-fact with at least a minimum of information for a brief profile of the project.

D.2.3 Relations with partners: Partnering internal and external to an organization

An NSF site visit team in 2011 held discussions with AtmChem. The team found AtmChem would benefit from creating a more 'viable' approach to its mission that involved a unique integrative blend of fieldwork and laboratory observations, modeling, and services being provided to university communities. In 2014, a change initiative focused on establishing a center in cooperation with NSF. Mission statement goals included building "a better alliance between AtmChem and University partners in order to address existing and emerging questions in observational atmospheric chemistry, and to provide community input regarding the role of a national center in the area of in situ atmospheric chemistry measurements". Paraphrasing the 2014 division report, an AtmCenter outcome was reported: "To meet the goals and objective outlined in the center's strategic plan, a laboratory was dissolved on March 1, 2015, and its three internationally respected research divisions became three laboratories at the center." AtmChem, one of the three laboratories, augmented its name with the addition of 'Observations and Modeling' to emphasize its unique features, the breadth of its scope, and the tight coordination between data collecting and computational modeling. During this period of AtmChem reorganization, an AtmDM participant expressed commented on their interactions with AtmChem:

> If we could somehow connect with other divisions and share the nuts and bolts of actually archiving and providing data that would be helpful. It seems the chemistry data is somewhat unique. And the way AtmDM is doing things and the way the climate models

> are distributed maybe aren't exactly what is most useful for AtmChem is sort of my feeling. (RS)

The notion that a unit's data 'is somewhat unique' represents a moment of discernment in the balance of local and general design features of information systems.

D.2.4 AtmCenter stories of participation

The following stories of participation and partnering occurred over the eight months of my fieldwork (2013- 2014). The first is a data management presentation made to AtmChem and the second a poster on organization-wide data efforts.

Participation story 1: Briefing series talk on data curation

In my work with AtmDM and AtmChem, I was invited to speak at a biweekly AtmChem Briefing Series Forum in the Spring of 2014. Because the AtmChem division of the organization did not include an individual trained in data management, I addressed two questions: 'Where is the AtmChem data?' and 'What is data curation?' (Baker, 2014a). The absence of formal discussion of these questions appeared to be one source of misunderstandings about data arrangements and data work. My presentation was unusual in considering the division's data collectively as a topic in their bi-weekly forum.

Discussion forums existed for individual projects while discussion about the division's management of the collection of many field projects appeared of less concern. Division-level data support did include a systems administration group effectively providing data storage and exchange, a web master providing web content delivery, and a research scientist with experience in creating 'merge files' who was oriented toward using data for research rather than data preservation. In professional presentations, AtmChem researchers typically spoke about environmental observations, modeling, and improving models with observations. Discussions of data from a single project sometimes occurred in researcher-led events known as data workshops. The discursive and deliberative workshops were described as forums that supported knowledge making via collective discussion of data, identification of data anomalies, and synthesis of information from selected sets of data.

To address "Where is the AtmChem data", I reported on results of my inventory in 2014 of the AtmChem data catalog posted online. The inventory revealed that data held in differing forms and locations was not distinguished on the division's website. In addition, a number of links in the online project data list were broken. The survey showed overall for the data list that 32% of the links were to the AtmDM file storage, 22% to NASA or NOAA archives, 20% required update, and 26% had no link given. Participants were not surprised by these results having experienced and accepted earlier that difficulties in data availability were the norm. As a contribution to vocabulary building, three elements of the AtmDM multi-project archive were reviewed: the field catalog, the project landing page, and the project dataset catalog. Each of these represented a different entry point into the data depending on whether a researcher was interested in a view of data created during field deployment, of the project as a whole together with its participants, or of a series of projects. Few researchers were clear on distinguishing the three. It takes time, use, and continuing dialogue to develop familiarity rather than impatience with the nuanced interconnected data entry points created by highly developed data systems.

To address "What is data curation", a two-stream model was presented (Baker et al., 2013). This model focused on two distinct streams of data activities: internal project use of data

and external or data reuse. Researchers were familiar with how data plays a central role in the production of knowledge that is reported in scientific publications. Project-specific use of data often involved a complex mix of processing, analysis, and integration strategies. External use of data, however, required data to be prepared with a more standardized set of procedures that created data packages in the form of well-described, measurement-based datasets for public access. Data use was the purview of the AtmDM group though was not generally recognized as new work required to support access, cross-project standardization, and data preservation.

Participation story 2: Data stewardship organization-wide poster

Amidst discussions of open data and national coordination across data repositories, there was also need for coordination of data work and repositories within the organization. At AtmCenterOrg, data collections began in the1960s with individual labs and divisions providing their own data tools and services. Thinking about organization-wide digital activities began around 2000 and was aimed at developing more coherence in data services across the center. These data activities were embedded within an evolving IT infrastructure with each coordination effort constrained by very different technologies and data practices. One participant recalled that IT coordination was the central concern:

> We recently had this IT summit (Feb 2013) of all of the IT people from the various entities at AtmCenter. They exchanged a lot of ideas about things that we should focus on in our infrastructure. But data seemed like a such a large problem that I don't think we really knew even how to get the conversation. (TS)

Data was a 'problem' that had not matured into a topic that could be addressed. Though the summit was informed by past efforts and spawned a subcommittee that fed into the AtmCenter strategic plan, software engineers found it difficult to identify how to proceed with internal efforts spanning all the divisions:

> I think it's hard for us to even articulate what the issue is. We all kind of think we know but to me its about safety of data. You know we just loose too much data to hardware failure. We lose data because metadata gets lost … we made poor choices about how to format data. Like [with] documents and ASCII word processors that we can't read anymore. (TS)

Four major organization-wide efforts occurring sequentially over almost two decades demonstrate that alignment is not a one-time event but an iterative, ongoing process of envisioning and enacting data work. A history of these organization-wide efforts was captured in a poster authored by three information scientists together with three software engineers who lead data efforts at the center (Baker et al., 2015c). Four distinct data initiatives were identified: in 1995 when Information Infrastructure Technology Applications was part of an information technology strategic plan, in 2001 when a Data Management Working Group created a community data portal, in 2010 when an ad hoc Committee on Data Citation coordinated Digital Object Identifier services, and in 2014 when a Data Stewardship and Engineering Team (DSET) was the first data group formally supported and recognized organization-wide. DSET's aim included creating a common discovery of resources. The poster prompted reflection about organization-wide data efforts. The evolution of an intra-organizational understanding of data –

beyond the individual research group or project – was considered critical to data management interoperability and capacity building. The poster served to document and communicate the process of organizing internal coordination.

APPENDIX E. AtmDM Data Work

When selected as the archive for a field project, the role of AtmDM was to support data efforts proposed and organized by researchers in AtmCenter and by members of the atmospheric university community. During the period 2005-2014, AtmDM supported approximately 8 to 15 projects per year using NCAR's observing platforms (e.g. aircraft) and instruments (e.g. atmospheric profilers).

AtmDM's mission statement was encapsulated as 'Four D's': Development, Deployment, Data Services, and Discovery. Data services provided during stages ranging from the fieldwork to the archive that included:

- Prefield planning – data policy, protocols, plan development, and composite-creation
- Data storage – data systems storing data that support version control
- Field catalog - a communication hub during field activities
- Project web pages – a hub for project related websites, publications, and artifacts
- Data processing – reformatting data, creating metadata, performing quality control
- Project dataset master list – list of the datasets associated with a project
- Dataset access – dataset metadata and data order forms (CODIAC)
- Project legacies - archive, publications, and metrics

A prefield planning meeting was key to launch of data management for a field campaign. At this time project data requirements as well as optional services were discussed. Requirements included creating a personalized data policy, addressing data format issues, identifying documentation including guidelines, considering project logistics including mailing lists, developing a field catalog as well as a project web page. In addition, the group made available climate 'composites' that brought together and integrated diverse data sources resulting in a digital preview of the area to be sampled. During fieldwork, situation awareness was maintained while assisting with instrumentation on aircraft, file storage, file transfer, and maintaining a field catalog. For post fieldwork, AtmDM managed data systems undergoing continuous redesign in responding to changes in technology, field support requirements, integration of system functionalities, and extra-organizational data repository interactions.

Table E.1 gives a summary of AtmDM in the same manner as for the three cases (Tables 4.1, 4.3, and 4.4) with a brief case description followed by my fieldwork and observations).

Table E.1: Overview of the AtmDM unit of AtmCenter

Case Name	Case Research	Case Description	Study Fieldwork	Study Activities Observed
AtmDM	Atmospheric field data	* AtmCenter planning begins in 1957 * AtmDM formed 2005 by merge of two units * Support for community field projects * Fieldwork global at multiple remote sites * Designated team of software engineers * Mature data archive services	2013-2014	* Support for multiple projects * Preserving project collections * data systems continuing design

In concert with other groups within their laboratory that focus on software systems, software applications, collaborative technologies and metadata, the group provided support before, during, and after field campaigns. Together with researchers, the AtmDM group of software engineers was participating constantly in field campaigns in addition preserving data from field projects. AtmDM provides an example of a centralized archive with data services within an organization that supports a variety of field projects each year.

APPENDIX F. LTER Information Committee History and Guides

F.1 LTER Information Management Development, A Brief History

The LTER approach to data management is illustrative of a slow growth approach to development of data systems and infrastructure fostered by small, steady budgets and continuing communication (Baker and Karasti, 2004; Karasti et al., 2006; Michener et al., 2011; Karasti et al., 2010; Jackson and Barbrow, 2015). Time and attention first and foremost are dedicated to data care guided originally by two aims: 1) to maximize access to data in support of ongoing research and 2) to minimize disturbance to research. In the mid 1990s, a third aim was added: 3) development of data sharing capabilities and policies (Porter, 2010). Slow growth allowed time for assessing and prioritizing local data needs while simultaneously coordinating and learning across sites via comparative analysis inherent to network activities. The LTER Information Management Committee, a standing committee with members from each site, dedicates time to community communication by generating reports on annual meetings, documenting best practices, and publishing a 'Databits' newsletter twice a year for a number of years with featured articles, commentary, news bits, good reads and good tools (LTER Databits, nd). Within this tradition of open communication, each LTER member project's information manager is also responsible for communications with all members of the local project.

The LTER Information Managers are informed and influenced by an array of data-related efforts led by ecological researchers, informatics specialists, and site-based information managers (e.g. Michener et al., 2011; Porter, 2010; Laney et al., 2013). A network-wide paper-based data catalog was developed in 1990 (Michener et al., 1990), a decade after the LTER began. It was followed by sites making data available on the internet since it eased access issues encountered with local computer logins by distributed teams. In addition, the internet enabled the LTER Network Office to develop a digital all-site personnel directory. Development of an LTER all-site bibliography provided experience with the diversity of site development capabilities and multiplicity of approaches to citation management (Chinn and Bledsoe, 1997).

As the digital era continued, new sites were added to the network and development of local data systems at each site became an established tradition. In addition, an LTER network information system (NIS) was envisioned, initially in the form of a set of modules (Brunt, 1998; Baker et al., 2000). Design and development of NIS modules was first imagined as the responsibility of site data managers. Site members led development of NIS modules such as a data catalog, a site description directory, key word vocabulary, and a climate data system (Brunt and Porter; 1999; Baker et al., 2002; Henshaw et al., 2006).

Projects of large scope at domain and national levels were typically led from outside the Information Management Committee by lead scientists or network partners working together with the LTER Network Office and information management working groups. An ecological metadata specification known as a 'non-geospatial metadata for the ecological sciences' was developed by an Ecological Society of America working group (Michener et al., 1997). In partnership with a national center, this developed into an XML template that was adopted by the LTER community and enacted within the network (Millerand and Bowker, 2009). A best practices guide was developed by the information managers for use of EML within their community. Local data systems matured and developed the capacity to deliver EML metadata

files while the LTER Network Office developed a harvester and Metacat, a central database for storing EML documents (Brunt, 2004).

Experience was growing within LTER with large-scale information system architectures at sites (e.g. Ellison et al., 2006) and the network with EcoTrends (Peters et al., 2013; Servilla et al, 2008) and a prototype architecture to automate the integration of time-series data dubbed Provenance Aware Synthesis Tracking Architecture (PASTA) (Servilla et al., 2006). The unanticipated availability of 'stimulus' funds in 2010 to design and develop NIS at the LTER Network Office prompted a shift from site-led NIS efforts. NIS initially assembled EML metadata file in XML format that linked to distributed data but eventually included centralization of data as well in order to avoid difficulties of broken links and unavailable systems (Michener et al., 2011). PASTA matured as a back-end service framework to manage data, metadata, and data packages (Servilla et al., 2016). Once ingested, a metadata and data congruency checker for quality assurance and correctness is activated and a report generated. A unique feature of the NIS and PASTA development was the incorporation of Tiger Teams that engaged information managers from the LTER project sites, thereby providing training for local personnel as well as bridging the design-use gap that accompanies development with many large-scale system. Such teams together with an LTER Information Management Committee working group joined together with system developers to develop system functionalities with particular focus on a dataset metadata checker (Servilla and Brunt, 2011; Bohm et al., 2012; Chamblee, 2013). The decade-long development of PASTA within the LTER community supports the LTER Data Portal (nd; Servilla et al., 2016).

F.2 LTER Information Management Guides

a. LTER Information Management Guides
b. Selected Books and Publications – LTER Information Management
c. Selected Books and Book Chapters – LTER Network
d. Selected Books and Works – Ecosystem Science and the International Biological Program

a. LTER Information Management Guides

The list of guides for information management include site functionality, practices and review criteria as well as on data access policies, spatial data documentation, web site design, terms of reference. In addition, there are best practices for metadata, units, controlled vocabularies, and documenting spatial data.

- Data Access Policy, 2005
  - http://www.lternet.edu/data/netpolicy.html
- LTER Information Management Practices, 2007
  - http://im.lternet.edu/resources/im_practices
- Critical Site Functionality, 2007
  - http://im.lternet.edu/im_requirements/critical_functionality
- Best Practices for Documenting Spatial Data, 2007
  - http://im.lternet.edu/im_practices/gis
- Review Criteria for LTER Information Management ,2009,
  - http://im.lternet.edu/im_requirements/im_review_criteria

- Guidelines for LTER Web Site Design and Content, 2009
  - http://im.lternet.edu/im_requirements/webdesign_guidelines
- Unit Best Practices, 2009
  - http://im.lternet.edu/sites/im.lternet.edu/files/LTERunitBestPractices_V12.doc
- Metadata Best Practices, 2011
  - http://im.lternet.edu/sites/im.lternet.edu/files/emlbestpractices-2.0-FINAL-20110801_0.pdf
- Terms of Reference, 2011
  - http://im.lternet.edu/im_requirements/TermsOfReference
- Controlled Vocabulary Best Practices, 2013
  - http://im.lternet.edu/VocabBestPractices
- Sensor Data Best Practices (moved to ESIP EnviroSensing Cluster)
  - http://wiki.esipfed.org/index.php/EnviroSensing_Cluster

b. Selected Books and Publications – LTER Information Management

Baker, K. S., & Millerand, F. (2010). Infrastructuring ecology: Challenges in achieving data sharing. *Collaboration in the New Life Sciences.* Ashgate.

Benson, B. J., Hanson, P. C., Chipman, J. W., & Bowser, C. J. (2006). Breaking the Data Barrier: Research Facilitation through Information Management. In J. J. Magnuson, T. K. Kratz & B. J. Benson (Eds.), *Long-Term Dynamics of Lakes in the Landscape: Long-Term Ecological Research on North Temperate Lakes*: Oxford University Press.

Karasti, H. & Baker, K. S. (2004) Infrastructuring for the long-term: Ecological information management. Proceedings of the 37th Annual Hawaii International Conference on System Science.

Karasti, H., Baker, K. S., & Halkola, E. (2006). Enriching the Notion of data curation in e-Science: Data managing and information infrastructuring in the Long Term Ecological Research (LTER) Network. Computer Supported Cooperative Work, 15(4), 321-358.

Michener, W. K., & Brunt, J. W. (2000). *Ecological Data: Design, Management and Processing*: Blackwell Science.

Michener, W. K., Porter, J., Servilla, M., & Vanderbilt, K. (2011). Long term ecological research and information management. *Ecological Informatics, 6*, 13-24.

Servilla, M., Brunt, J., Costa, D., McGann, J., & Waide, R. (2016). The contribution and reuse of LTER data in the Provenance Aware Synthesis Tracking Architecture (PASTA) data repository. *Ecological Informatics*. doi: doi:10.1016/j.ecoinf.2016.07.003

c. Selected Books and Book Chapters – LTER Network

Likens, G. E. (Ed.). (1987). *Long-Term Studies in Ecology: Approaches and Alternatives [Hardcover]*: Springer.

Golley, F. B. (1993). *A History of the Ecosystem Concept in Ecology: More than the Sum of the Parts*: Yale University Press.

Coleman, D. C. (2010). *Big Ecology: The Emergence of Ecosystem Science*: University of California Press.

Gosz, J. R., Waide, R. B., & Magnuson, J. J. (2010). Twenty-eight years of the US-LTER program: experience, results, and research questions. In F. Muller & C. Bassler (Eds.),

Long-term ecological research (Vol. Long-Term Ecological Research, Between Theory and Application, pp. 59-74): Springer.
Willig, M. R., & Walker, L. R. (Eds.). (2016). *Long-Term Ecological Research: Changing the Nature of Scientists*. New York: Oxford University Press.

d. Selected Books and Works – Ecosystem Science and the International Biological Program

Aronova, E., Baker, K. S., & Oreskes, N. (2010). Big science and big data in biology: From the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) network, 1957–present.
Kwa, C. (1987). Representations of nature mediating between ecology and science policy: the case of the International Biological Programme. *Social STudies of Science, 17*(3), 413-442.
Kwa, C. (1993). Modeling the grasslands. Historical studies in the physical and biological sciences, 24(1), 125-155.
Toby A. Appel, Shaping Biology: The National Science Foundation and American Biological Research, 1945–1975 (Baltimore: Johns Hopkins University Press, 2000);
Sharon E. Kingsland, The Evolution of American Ecology, 1890–2000 (Baltimore: Johns Hopkins University Press, 2005).
Conrad Waddington, "The Origin," in The Evolution of IBP, ed. E. B. Worthington (Cambridge: Cambridge University Press, 1975), 4–11.

APPENDIX G. Data Management Position Advertisements

G.1 EcoPrairie Job Advertisement: Colorado State University 2015

Data Management Specialist

Below you will find the details for the position including any supplementary documentation and questions you should review before applying for the opening. To apply for the position, please click the Apply for this Job link/button.

To apply, upload a letter of application addressing your qualifications for the position, a current resume or CV, plus names and contact information for three professional references. References will not be contacted without prior notification of candidates. In the cover letter, please address each of the required qualifications and as many preferred qualifications as are applicable. Note that when addressing knowledge of data-driven instruction, the term is used to determine the candidate's knowledge of how researchers learn to manage data; what faculty on campus are teaching students about gathering, processing, and analyzing data; and other aspects of data management such as structure and metadata.

Posting Details
Posting Detail Information
Working Title Data Management Specialist
Posting Number 201500244F
Proposed Annual Salary Range 57,000-65,000
Position Type Faculty
Work Hours/Week 40

Description of Work Unit
The Colorado State University Libraries is a technologically sophisticated environment with extensive electronic resources and service primarily located in Morgan Library. The Dean of Libraries is the campus Vice President for Information Technology, and Academic Computing and Network Services is part of the Libraries. CSU is making a strong and strategic investment to ensure that our academic libraries meet the needs of the institution, now and into the future. A strong collaboration exists and is burgeoning with the Vice President for Research. Digital Collection Services within CSU Libraries supports CSU's academic and research needs by providing data management services to researchers, faculty and students; creating and providing access to digital resources through digitization and the application of metadata; and managing the consortial and CSU digital repository, Digital Collections of Colorado (https://dspace.library.colostate.edu/) and its content. This position is responsible for the coordination and deployment of data services during the research lifecycle. This includes advising researchers on creating and implementing data management plans at CSU; partnering with the College Liaison Librarians in outreach to CSU faculty, working with the Data Management Committee, the Information Science & Technology Center (ISTeC), and the Office of the Vice President for Research; and providing training and consulting services. This position

offers an exciting opportunity to support a scholarly communications program that will provide support services relating to organizing, preserving, and sharing of research data.

Tenure Track? No
% Research 0
% Teaching 100
% Service 0

Open Posting Date
Open Until Filled Yes
To ensure full consideration, applications must be received by 11:59pm (MT) on 01/31/2016
Number of Vacancies 1
Desired Start Date
Position End Date

Position Summary
The Data Management Specialist guides the implementation of data management and curation services at CSU, and is responsible for coordinating, managing, preserving, and providing access to research data created by CSU researchers according to best practices and policies. The position promotes awareness of resources and services among faculty and students through comprehensive, customized instructional activities to faculty and students in academic departments and research units. The position also works closely with the staff of the Digital Library and ePublishing Services Division and other CSU Libraries units to ensure that acquired data resources are properly formatted, accessible, and discoverable. The position participates in strategic planning in the context of research data, research data management, and associated topics. Specific projects and outreach activities may require some evening and weekend work.

Required Job Qualifications
• Master's degree in library science from an ALA-accredited program or international equivalent by December 31, 2015; or graduate degree in a data-intensive field by December 31, 2015.
• Minimum of two years of experience in the development, coordination, delivery, and promotion of research or data management services.
• Minimum of two years of experience with tools for managing digital assets and data, such as DMPTool, Dataverse, and similar services.
• Experience with institutional or subject-based repository systems such as DSpace, Fedora, CONTENTdm or similar systems.
• Knowledge of data description and metadata standards such as Dublin Core, Ecological Metadata Language (EML), MODS, and XML; identifier schema such as DOI or ORCID; relevant content transmission and retrieval protocols, such as OAI, and others.
• Knowledge of data-driven instruction, research, scholarly communication, and the academic, technical and social aspects of data management.
• Knowledge of issues and technical challenges related to the life cycle of research data, including storage and preservation.

Preferred Job Qualifications
• Academic background in a science discipline.

• Experience in identifying researcher information needs and in creating effective services to meet those needs.
• Experience with developing and delivering workshops and training sessions.
• Experience with data analytics.
• Demonstrated knowledge of research data curation trends, including disciplinary and institutional repositories, and funder and publisher mandates.
• Knowledge of digital preservation standards and best practices.
• Ability to work independently and collaboratively within an exciting and rapidly evolving environment; ability to work well with a diverse employee and user community; strong analytical and decision-making skills; excellent interpersonal and communication skills.

Special Instruction to Applicants
To apply, upload a letter of application addressing your qualifications for the position, a current resume or CV, plus names and contact information for three professional references. References will not be contacted without prior notification of candidates. In the cover letter, please address each of the required qualifications and as many preferred qualifications as are applicable. Note that when addressing knowledge of data-driven instruction, the term is used to determine the candidate's knowledge of how researchers learn to manage data; what faculty on campus are teaching students about gathering, processing, and analyzing data; and other aspects of data management such as structure and metadata.

Background Check Policy Statement
Colorado State University (CSU) strives to provide a safe study, work, and living environment for its faculty, staff, volunteers and students. To support this environment and comply with applicable laws and regulations, CSU conducts background checks. The type of background check conducted varies by position and can include, but is not limited to, criminal (felony and misdemeanor) history, sex offender registry, motor vehicle history, financial history, and/or education verification. Background checks will be conducted when required by law or contract and when, in the discretion of the university, it is reasonable and prudent to do so.

EEO Statement
Colorado State University is committed to providing an environment that is free from discrimination and harassment based on race, age, creed, color, religion, national origin or ancestry, sex, gender, disability, veteran status, genetic information, sexual orientation, gender identity or expression, or pregnancy. Colorado State University is an equal opportunity/equal access/affirmative action employer fully committed to achieving a diverse workforce and complies with all Federal and Colorado State laws, regulations, and executive orders regarding non-discrimination and affirmative action. The Office of Equal Opportunity is located in 101 Student Services.

Diversity Statement
Reflecting departmental and institutional values, candidates are expected to have the ability to advance the Libraries' commitment to diversity and inclusion.

Search Contact Sande Innes, sande.innes@colostate.edu
Quick Link For Posting http://jobs.colostate.edu:80/postings/29811

Essential Job Duties
Essential Job Duties

Job Duty Category Data Management Consultation and Services
Duty/Responsibility
Performs data management planning with principal investigators and researchers prior to grant submissions, or before the research data is collected or created. Works with researchers to identify, recruit, ingest, and deposit data in the CSU digital repository, adhering to local policies and national and international standards and best practices, or in an appropriate subject repository.
Percentage Of Time 60

Job Duty Category Education and Community Building
Duty/Responsibility
Serves as a resource to researchers, librarians, and the university community on data management issues and services. Develops and delivers training on data curation, provides guidance and instruction on discovery, acquisition, and use of research data. Works with researchers including faculty, graduate and post-doc students, academic and administrative units and research centers to enable them to better manage, archive and make available their research data.
Percentage Of Time 30

Job Duty Category Data Management Policy and Strategy Development
In cooperation with supervisor, other members of the Libraries and university partners,
Duty/Responsibility
Contributes to planning, development, and evaluation of data management services to CSU faculty, researchers, and students. Develops roadmap to determine capacities and expertise needed to plan, implement, and evaluate a sustainable data curation service across the university.
Percentage Of Time 10

Supplemental Questions
Required fields are indicated with an asterisk (*).
1. * Do you/will you have a Master's degree in library science from an ALA-accredited program or international equivalent by Dec. 31, 2015; or graduate degree in a data-intensive field by Dec. 31, 2015?
Yes / No
2. * Do you have a minimum of two years of experience in the development, coordination, delivery, and promotion of research or data management services?
Yes / No
3. * Do you have a minimum of two years of experience with tools for managing digital assets and data, such as DMPTool, Dataverse, and similar services?
Yes / No
4. * Do you have experience with institutional or subject-based repository systems such as DSpace, Fedora, CONTENTdm, or similar systems?
Yes / No

5. * Do you have knowledge of data description and metadata standards such as Dublin Core, Ecological Metadata Language (EML), MODS, and XML; identifier schema such as DOI or ORCID; relevant content transmission and retrieval protocols, such as OAI, and others?
Yes / No
6. * Do you have knowledge of data-driven instruction, research, scholarly communication, and the academic, technical, and social aspects of data management?
Yes / No
7. * Do you have knowledge of issues and technical challenges related to the life cycle of research data, including storage and preservation?
Yes / No

G.2 EcoPrairie Job Advertisement: USDA 2015

Data Management Position

SALARY RANGE:	$58,562.00 to $76,131.00 / Per Year
OPEN PERIOD:	Tuesday, August 11, 2015 to Monday, August 17, 2015
SERIES & GRADE:	GS-0401-11
POSITION INFORMATION:	Full Time - Permanent
PROMOTION POTENTIAL:	11
DUTY LOCATIONS:	1 vacancy in the following location: Fort Collins, CO
WHO MAY APPLY:	United States Citizens
SECURITY CLEARANCE:	Q - Nonsensitive
SUPERVISORY STATUS:	No

JOB SUMMARY:
Find Solutions to Agricultural Problems that Affect Americans Every Day, From Field to Table
The mission of the Rangeland Resources Research unit (RRRU), headquartered in Cheyenne, Wyoming at the High Plains Grasslands Research Station, with site locations at Nunn, CO (Central Plains Experimental Range) and Fortt Collins, CO (Crops Research Laboratory), is to conduct research within two major research areas: 1) improving management to balance production and conservation goals in rangelands; and 2) ecosystem response to and management adaptation for a changing climate and weather variability, including extreme events. The position is located in the RRRU at the Fort Collins, CO location and the incumbent serves as a support scientist responsible for the data management activities involved in the Long-Term Agro-ecosystem Research (LTAR) network projects.

Work may require intensive periods of time working with computers, data and reports. Field work is required to observe data collection, methodology and protocols. In addition, work requires moderate physical exertion, sometimes requiring lifting heavy objects (up to 50 pounds) using appropriate weight bearing equipment and standing for prolonged periods of time. Travel by vehicle is required to field sites and further travel may be required to attend meetings.

TRAVEL REQUIRED
- Occasional Travel
- Less than 10 days per year

RELOCATION AUTHORIZED
- No

KEY REQUIREMENTS
- You must be a US Citizen or US National
- Males born after 12-31-59 must be registered for Selective Svc., or exempt
- Subject to favorable adjudication of a background investigation
- Successful completion of a one-year probationary or trial period
- Driver's License is required

DUTIES:
- Overseeing data management for the Rangeland Resources Research Unit (Cheyenne, Wyoming/Fort Collins. Colorado) including the Long-Term Agro-ecosystem Research (LTAR) network site at the Central Plains Experimental Range (CPER) at Nunn, Colorado.
- Consulting with the supervisor to plan and determine approaches to assignment concerning LTAR projects at the CPER as well as cross-site LTAR projects focused on an associated set of complementary objectives within an overall national LTAR mission
- Designing, maintaining, documenting and monitoring processes and workflow related to data acquisition, maintenance, organization, accessibility, and integrity
- Ensuring the maintenance and stewardship of datasets describing rangeland plants, animals, soils, climate, land use legacies and management strategies, which includes existing relational databases and Geographic Information Systems (GIS)/Global Positioning System (GPS) data sets and projects, continuous climatic and gas flux exchange data sets, and long-term data sets with shifting methods as well as construction of new databases and user input interfaces
- Implementing QA/QC (quality assurance/quality control) standards, oversight of data transfer and documentation protocols, population of data dictionaries, implementation of data certification procedures, and maintenance of archival and retrieval protocols
- Coordination of the RRRU's database management with other LTAR network sites to facilitate regional and national cross-site data analysis and synthesis.

QUALIFICATIONS REQUIRED:

Basic Requirement:
Successful completion of a full 4-year course of study in an accredited college or university leading to a bachelor's or higher degree that included a major field of study or specific course requirements in biological sciences, agriculture, natural resource management, chemistry, or related disciplines appropriate to the position. The number of semester hours required to constitute a major field of study is the amount specified by the college or university attended. If this number cannot be obtained, 24 semester hours will be considered as equivalent to a major field of study. The nature and quality of this required course work must have been such that it would serve as a prerequisite for more advanced study in the field or subject-matter area. Related

course work generally refers to courses that may be accepted as part of the program major. (MUST SUBMIT TRANSCRIPTS)
OR
Combination of education and experience -- Courses equivalent to a major, as shown above, plus appropriate experience or additional education. The quality of the combination of education and experience must be sufficient to demonstrate that the applicant possesses the knowledge, skills, and abilities required to perform work as a Biologist, and is comparable to that normally acquired through the successful completion of a full 4-year course of study with a major in the appropriate field. In addition to courses in the major and related fields, a typical college degree would have included courses that involved analysis, writing, critical thinking, research, etc. These courses would have provided an applicant with skills and abilities sufficient to perform progressively more responsible work in the occupation. Therefore, creditable experience should have demonstrated similarly appropriate skills or abilities needed to perform the work of the occupation. (MUST SUBMIT TRANSCRIPTS)

In addition to meeting the basic requirement, you must also meet one of the minimum qualifications below.
Minimum Qualifications at the GS-11 Level:
Specialized Experience: At least one full-time year (12-months) of specialized work experience equivalent to the GS-09 grade level in the Federal service. Examples of specialized experience include but are not limited to:

- Coordinating management of biological, agricultural, ecological, or natural resource data from multiple sites to facilitate the delivery of data among scientists;
- Evaluating needs for management, acquisition, and enhancement of biological, agricultural, ecological, or natural resource data;
- Updating and maintaining research databases;
- writing laboratory reports to explain experimental design, principle, procedure, statistical analysis, and results; and
- performing computerized data collection, statistical analyses, and data mining.

OR
Education: Successful completion of 3 years of progressively higher level graduate education leading to a Ph.D. degree or a Ph.D. or equivalent doctoral degree in quantitative biology, computational biology, bioinformatics, biomathematics, or other closely related field which demonstrates the knowledge, skills and abilities required by this position. Note: One year of full-time graduate education is considered to be the number of credit hours that the school attended has determined to represent 1 year of full-time study. If that information cannot be obtained from the school, 18 semester hours should be considered as 1 year of full-time study. (MUST SUBMIT TRANSCRIPTS)
OR
Combination: Applicants may qualify based on an appropriate combination of successfully completed graduate education and experience. To determine your combination, first compute your experience as a percentage of the experience listed above; then determine your graduate education as a percentage of the education listed above; then add the two percentages. The total percentages must equal at least 100 percent. Only graduate education in excess of two years (36 semester hours) may be used to qualify. (MUST SUBMIT TRANSCRIPTS)

IN DESCRIBING YOUR EXPERIENCE, PLEASE BE CLEAR AND SPECIFIC. WE MAY NOT MAKE ASSUMPTIONS REGARDING YOUR EXPERIENCE. If your resume does not support your questionnaire answers, we will not allow credit for your response(s).
For more information on the qualifications for this position, click here:
http://www.opm.gov/qualifications/Standards/group-stds/gs-admin.asp

HOW YOU WILL BE EVALUATED:
Applications will be evaluated in accordance with Office of Personnel Management's (OPM) Delegated Examining Procedures using category rating. Applicants who meet basic minimum qualifications will be placed in one of three categories: Best Qualified, Well Qualified, or Qualified. Within these categories, applicants eligible for veteran's preference will receive selection priority over non-veterans. Category placement will be determined based on applicants' quality of experience and the extent they possess the following knowledge, skills, and abilities (or competencies):

· Knowledge of the experimental methods and protocols for data acquisition for biological, agricultural, ecological, or natural resource research projects
· Knowledge of the theories, principles and practices of information and data management
· Skill in the design and management of a database
· Skill in using computer programming and software programs for data mining and to create, manage, and share datasets
· Ability to communicate orally
· Ability to communicate in writing

The questionnaire will assess your qualifications for the job, and will be used to identify the best qualified applicants to be referred to the hiring manager for further consideration and possible interviews. Your ratings in this Assessment Questionnaire are subject to evaluation and verification based on the documents and references you submit. Later steps in the selection process are specifically designed to verify your ratings. Attempts to falsify information; inflate your qualifications or providing inaccurate information on federal documents may be grounds to adjust your rating or to not select you. Errors, omissions or providing inaccurate information on federal documents may affect your eligibility. If selected providing inaccurate information on federal documents could also be grounds for dismissing you from the position/agency. Please follow all instructions carefully.

G.3 EcoRiver Job Advertisement: The Nature Conservancy 2016

Data Management Position

JOB TITLE: Conservation Information Manager III
JOB FAMILY: Conservation
JOB NUMBER: XX
SALARY GRADE: X
STATUS: Salaried, 12 months, extension pending
DATE: xxxxx

SUMMARY:
The Conservation Information Manager III designs, manages, maintains, and delivers conservation data and provides support for data activities including work with technologies, data systems, data sharing and documentation to Conservancy staff.

ESSENTIAL FUNCTIONS:
* Participates in planning of data & information environments supporting conservation at multiple levels.
* S/he performs data management and data processing; familiar with differing information environments designs and data assembly, produces professional reports, processes data sets and carries out data tasks relating to tabular source material and advanced queries, and provides hardware/software support.
* Develops and delivers training to staff individually and in groups, produces maps and other graphic products and reports.
* Designs, documents, maintains, and monitors processes and workflow related to data acquisition, maintenance, organization, accessibility, and integrity.
* Implements QA/QC (quality assurance/quality control) standards and oversees data transfer and documentation protocols including documentation of data collection methods, population of data dictionaries, and implementation of data certification procedures.
* Ensures long-term stewardship of datasets by developing & maintaining archival & retrieval protocols.
*Establishes procedures for data accessibility, sharing, and security.

RESPONSIBILITIES & SCOPE
* Supervises staff and has ability to motivate, lead, set objectives and manage performance, including conflict resolution.
* May help develop and manage data work plans and large project budgets.
* May negotiate and contract with vendors.
* Ensures integrity of data collecting & management relating to conservation projects & project statistics.
* Acts independently, under limited supervision, resolves complex issues within the program area and may act as a resource to others.

MINIMUM QUALIFICATIONS
* MS or BS degree and certification in a related field and 3 years related experience, or equivalent combination of education and experience. Related fields include biological sciences, agriculture, natural resource management, chemistry, or related disciplines, such as information sciences appropriate to the position. The nature and quality of this required course work must have been such that it would serve as a prerequisite for more advanced study in the field or subject-matter area.
* Experience managing, maintaining and populating databases and manual files.
* Experience with image analysis and interpretation, complex spatial analysis, data modeling and landscape scenario analysis.
* Experience managing multiple projects.
* Experience operating GIS software, analyzing data and producing data reports and creating maps.

* Experience in work collaboratively and developing/delivering training to practitioners.
* Experience with Microsoft Word, Excel, Access and delivery of content via the Web.

REFERENCES

Aguilar, R., Pan, J., Gries, C., San Gil, I., & Palanisamy, G. (2010). A flexible online metadata editing and management system. *Ecological Informatics, 5*(1), 26-31.

Aknan, A., Chen, G., Crawford, J., & Williams, E. (2013). *ICARTT file format standards V1.1 standards track*: Technical Report ESDS-RFC-019v1.1, National Aeronautics and Space Administration.

Allison, L., & Gurney, R. (2014). Belmont forum e-infrastructures and data management collaborative research action. Available at http://bfe-inf.org/documents.

Amit, V. (2000). Introduction: Constructing the field. In V. Amit (Ed.), *Constructing the Field: Ethnographic Fieldwork in the Contemporary World* (pp. 1-18). London: Routledge.

ARL (2006). To stand the test of time: Long-term stewardship of digital data sets in science and engineering. *American Research Libraries*. Arlington, VA.

ARL (2009). The research library's role in digital repository services. Final Report of the ARL Digital repository issues task force: Association of Research Libraries. Available at http://www.arl.org/storage/documents/publications/repository-services-report-jan09.pdf.

Armbruster, C., & Romary, L. (2010). Comparing repository types: Challenges and barriers for subject-based repositories, research repositories, national repository systems and institutional repositories in serving scholarly communication. Available at SSRN http://ssrn.com/abstract=1506905.

Aronova, E., Baker, K. S., & Oreskes, N. (2010). Big science and big data in biology: From the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) network, 1957–present. *Historical Studies in the Natural Sciences*, 40(2), 183-224.

AtmChemPC (nd). AtmChem project catalog of field campaigns. Available at https://www2.acom.ucar.edu/campaigns.

AtmDMPC (nd). AtmDM project catalog. Available at https://www.eol.ucar.edu/all-field-projects-and-deployments.

Baker, K.S. (2005). Informatics and the environmental sciences. Scripps Institution of Oceanography Technical Report, June 2005. Available at http://escholarship.org/uc/item/0179n650.

Baker, K. S. (2014a). Data curation, data production, and a web of repositories. *Presentation at bi-weekly research report meetings of Atmospheric Chemistry Divison, April 15, 2014.* National Center for Atmospheric Sciences, Boulder, CO.

Baker, K. S. (2014b). *Developing a data management strategy: Infrastructure for the Emiquon partnership. Talk.* Paper presented at the Emiquon Annual Science Conference 2014. 27 March 2014, Illinois State Museum, Dickson Mounds, Lewiston, IL.

Baker, K.S. (2016). Therkildsen field station at Emiquon: Data management beginning in 2013. Center for Informatics Science and Scholarship (CIRSS). The University of Illinois Urbana-Champaign. Available at http://hdl.handle.net/2142/94951.

Baker, K. S., Benson, B. J., Henshaw, D. L., Blodgett, D., Porter, J. H., & Stafford, S. G. (2000). Evolution of a multisite network information system: The LTER information management paradigm. *BioScience, 50*(11), 963-978.

Baker, K. S., Brunt, J. W., & Blankman, D. (2002). Organizational informatics: site description directories for research networks. Proceedings of the 6th World Multi-Conference on

Systematics, Cybernetics, and Informatics, Orlando.
Baker, K. S., & Duerr, R. E. (2016a). Data and a diversity of repositories. In L. Johnston (Ed.), *Curating Research Data: A Handbook of Current Practice* (Vol. 2): Association of College and Research Libraries.
Baker, K. S., & Duerr, R. E. (2016b). Research and the changing nature of data repositories. In L. Johnston (Ed.), *Curating Research Data: Practical Strategies for Your Digital Repository* (Vol. 1): Association of College and Research Libraries.
Baker, K. S., Duerr, R. E., & Parsons, M. A. (2015a). Scientific knowledge mobilization: Co-evolution of data products and designated communities. *International Journal of Digital Curation, 10*(2), 110-135.
Baker, K. S., Kaplan, N., & Melendez-Colom, E. (2010). IMC Governance working group: Developing a terms of reference. *LTER Databits Newsletter, Fall 2010*. Available at http://databits.lternet.edu/fall-2010/imc-governance-working-group-developing-terms-reference.
Baker, K. S., & Karasti, H. (2004). *The long-term information management trajectory: Working to support data, science and technology*: Scripps Institution of Oceanography, http://escholarship.org/uc/item/7d64x0bd.pdf.
Baker, K.S., Mayernik, M.S., Thompson, C.A., Nienhouse, E., Williams, S., & Worley, S. (2015c). Envisioning and enacting a coherent organization-wide view of data. International Conference for Digital Curation 2015, London.
Baker, K. S., & Millerand, F. (2007). Articulation work supporting information infrastructure design: coordination, categorization, and assessment in practice. 40th Annual Hawaii International Conference on System Sciences, HICSS 2007.
Baker, K. S., & Millerand, F. (2010). Infrastructuring ecology: Challenges in achieving data sharing. In B. Penders, J. N. Parker & N. Vermeulen (Eds.), *Collaboration in the New Life Sciences*: Ashgate Publishing, Ltd.
Baker, K. S., Millerand, F., & Yarmey, L. (2009). *Growing Information Infrastructure: Data lifecycles and subcycles.* Paper presented at the Poster Number 310. LTER ASM 2009: Long Term Ecological Research All Scientists Meeting, Sept 14-16, 2009. Available at http://asm.lternet.edu.
Baker, K. S., Palmer, C. L., Thomer, A. K., Wickett, K., DiLauro, T., Asangba, A. E., … Choudhury, G. S. (2013). *Two-stream model: Toward data production for sharing field science data.* Paper presented at the 46th Annual Fall Meeting of the American Geophysical Union 2013, San Francisco, CA.
Baker, K. S., Troxell-Thomas, C. A., & Pooler, W. G. (2015b). Data management: File systems, databases, and metadata. Poster. Poster: *Emiquon Annual Science Conference 2015, 19 Feb 2015* Dickson Mounds Museum, Lewistown, IL. Available at http://hdl.handle.net/2142/73366.
Baker, K. S., & Yarmey, L. (2009). Data stewardship: Environmental data curation and a web-of-repositories. *International Journal of Digital Curation, 4*(2), 12-27.
Ball, A. (2012). Review of the state of the art of the digital curation of research data. Bath, UK: University of Bath. Available at http://opus.bath.ac.uk/18774/2/erim1rep091103ab11.pdf.
Bannon, L. (1991). From human factors to human actors: The role of psychology and human-computer interaction studies in system design. *Design at work: Cooperative design of computer systems*, 25-44.
Bannon, L., & Schmidt. K. (1991) CSCW: Four characters in search of a context. *Studies in*

Computer Supported Cooperative Work: Theory, Practice and Design. Amsterdam: North Holland, 3-16.

Bannon, L., & Bødker, S. (1997). *Constructing common information spaces.* Paper presented at the Proceedings of the Fifth European Conference on Computer Supported Cooperative Work.

Bateson, G. (2000). *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*: University Of Chicago Press.

Beachy, R. (2010). Research at the USDA: Addressing societal grand challenges. Available at https://www.aaas.org/sites/default/files/migrate/uploads/RogerBeachy_AAASForum2010.pdf

Beagrie, N., Chruszcz, J., & Lavoie, B. (2008), Keeping research data safe: A cost model and guidance for UK universities, Final Report April 2008. Available at http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.doc.

Beaulieu, A. (2010). From co-location to co-presence: Shifts in the use of ethnography for the study of knowledge. *Social Studies of Science, 40*(5), 453-470.

Becker, H. S. (1970). *Sociological work: Method and substance*. New Brunswick, NJ: Transaction Books.

Benson, B. J., Hanson, P. C., Chipman, J. W., & Bowser, C. J. (2006). Breaking the data barrier: Research facilitation through information management. In J. J. Magnuson, T. K. Kratz & B. J. Benson (Eds.), *Long-Term Dynamics of Lakes in the Landscape: Long-Term Ecological Research on North Temperate Lakes*. Oxford University Press.

Berente, N. (2009). *Institutional logics and loosely coupled practices: The case of NASA's enterprise information system implementation.* (PhD), Case Western Reserve University. Bergold and Thomas, 2012

Bergold, J., & Thomas, S. (2012). Participatory research methods: A methodological approach in motion. *Historical Social Research/Historische Sozialforschung*, 191-222.

Berman, F., & Cerf, V. (2013). Who will pay for public access to research data?. *Science, 341*(6146), 616-617.

Berman, F. (2014). Building global infrastructure for data sharing and exchange through the research data alliance. D-Lib, *20*(1/2).

Billick, I. (2010). Managing place-based data, The Rocky Mountain Biological Laboratory as a case study, Chapter 17. In I. Billick & M. V. Price (Eds.), *The Ecology of Place: Contributions of Place-Based Research to Ecological Understanding*. Chicago: University of Chicago Press.

Birnholtz, J. P., & Bietz, M. J. (2003). *Data at work: Supporting sharing in science and engineering.* Paper presented at the Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work.

Bisco, R. L. (1967). Social science data archives: progress and prospects. *Social Science Information, 6*, 39-74.

Blomberg, J., & Karasti, H. (2013a). Reflections on 25 years of ethnography in CSCW. *Computer Supported Cooperative Work (CSCW)*(22), 373-423.

Blomberg, J., & Karasti, H. (2013b). Ethnography: Positioning ethnography within participatory design. *Routledge Handbook of Participatory Design*, 86.

Blomberg, J., Giacomi, J., Mosher, A., & Swenton-Wall, P. (1993). Ethnographic field methods and their relation to design. In D. Schuler & A. Namioka (Eds.), *Participatory Design: Principles and Practices* (pp. 123-155). Hillsdale, NJ: Erbaum Associates.

Bocking, S. (1990). Stephen Forbes, Jacob Reighard, and the emergence of aquatic ecology in the Great Lakes region. *Journal of the History of Biology, 23*(3), 461-498.

Boden, A., Rosswog, F., Stevens, G., & Wulf, V. (2014). *Articulation spaces: Bridging the gap between formal and informal coordination.* Paper presented at the Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing.

Bohm, S., Boose, E., Costa, D., Downing, J., Gastil-Buhl, M., Gries, C., … Servilla, M. (2012). Report to the IMC, EML data package checks and the PASTA quality engine. Long Term Ecological Research, Information Management Committee.

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology, 63*(6), 1059-1078.

Borgman, C. L. (2015). Big data, little data, no data: Scholarship in the networked world. Cambridge MA: MIT Press.

Bose, R., & Frew, J. (2005). Lineage retrieval for scientific data processing: A survey. *ACM Computing Surveys (CSUR), 37*(1), 1-28.

Bossen, C. (2002). The parameters of common information spaces: The heterogeneity of cooperative work at a hospital ward. Paper presented at the Proceedings of the 2002 ACM conference on Computer supported cooperative work.

Botero, A. & Saad-Sulonen, J. (2010). Enhancing citizenship: The role of in-between infrastructures. In PDC 2010: Proceedings of the 11th Biennial Participatory Design Conference. New York: ACM Press, pp. 81–90.

Bowen, G. M., & Roth, W.-M. (2002). The "socialization" and enculturation of ecologists in formal and informal settings. *Electronic Journal of Science Education, 6*(3).

Bowker, G. C., Baker, K., Millerand, F., & Ribes, D. (2010). Towards information infrastructure studies: Ways of knowing in a networked environment. In J. Hunsinger, L. Klastrup & M. Allen (Eds.), *International Handbook of Internet Research* (pp. 20). New York: Springer.

Bowker, G., Star, S. L., Gasser, L., & Turner, W. (Eds.). (1997). *Social Science, Technical Systems, and Cooperative Work: Beyond the Great Divide. Computers, Cognition and Work Series,* Psychology Press.

Box (nd). Box.com. Available at https://www.box.com/home.

Brase, J., Lautenschlager, M., & Sens, I. (2015). The tenth anniversary of assigning DOI names to scientific data and a five year history of DataCite. *D-Lib Magazine, 21*(1).

Brunt, J. W. (1998). The LTER network information system: A framework for ecological information management. In Aguirre-Braveo & C. R. Franco (Eds.), *North American Science Symposium: Toward a Unified Framework for Inventorying and Monitoring Forest Ecosystem Resources*, 2-6 Nov 1998, Guadalajara, Mexico. Fort Collins (CO): US Department of Agriculture, Forest Service, Rocky Mountain Research Station Proceedings RMRS-P-12.

Brunt, J. (2004). It was an exciting summer for informatics in the LTER Network and here are few highlights. *LTER Network News, Fall 2004*. Available at http://news.lternet.edu/Article386.html.

Brunt, J. W., & Porter, J. H. (1999). Ecological Metadata—in Perspective. *The Network Newsletter, 12*(2), 9-10. Available at http://news.lternet.edu/Article984.html.

Buckland, M. (1999). *Vocabulary as a central concept in library and information science.* Paper presented at the Proceedings of the Third International Conference on Conceptions of Library and Information Science (CoLIS3), 23-26 May 1999, Dubrovnik, Croatia.

Callahan, J. T. (1984). Long-term ecological research. *BioScience, 34*(6), 363-367.
Carlson, J. (2014). The use of life cycle models in developing and supporting data services. *Research Data Management: Practical Strategies for Information Professionals* (pp. 63-86): Purdue University Press.
Carroll, J. M. (2007). Learning in communities: Introduction to the special issue. *Computer Supported Cooperative Work* (16), 373-374.
Carroll, K. (2014). Body dirt or liquid gold? How the 'safety' of donated breastmilk is constructed for use in neonatal intensive care. *Social Studies of Science, 44*(3), 466-485.
CASRAI (nd). Consortia Advancing Standards in Research Administration Information Dictionary. Available at http://dictionary.casrai.org.
CCSDS (2012). Consultative committee for space data systems, reference model for an Open Archival Information System (OAIS). Washington DC: CCSDS 650.0-M-2, Magenta Book. Issue 2. June 2012. Available at http://public.ccsds.org/publications/archive/650x0m2.pdf.
Chamblee, J. (2013). The IMC meets the PASTA challenge. *LTER Databits Spring 2013*. Available at http://databits.lternet.edu/spring-2013/imc-meets-pasta-challenge.
Chamblee, J., & O'Brien, M. (2013). The IMC meet the PASTA challenge. *LTER Databits, Information Management Newsletter of the Long Term Ecological Research Newsletter*. Available at http://databits.lternet.edu/spring-2013/imc-meets-pasta-challenge
Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*: Sage Publications Limited.
Chinn, H., & Bledsoe, C. (1997). Internet access to ecological information-the US LTER All-Site Bibliography Project. *BioScience, 47*(1), 50-57.
Clarke, A. E. (1991). Social worlds/arenas theory as organizational theory. In D. R. Maines (Ed.), *Social Organization and Social Process: Essays in Honor of Anselm Strauss* (pp. 119-158). New York: Aldine de Gruyter.
Clarke, A. (2005). *Situational Analysis: Grounded Theory After the Postmodern Turn*: Sage.
Clarke, A. (2009). From grounded theory to situational analysis: What's new? Why? How? In J. M. Morse, P. N. Stern, J. Corbin, B. Bowers, K. Charmaz & A. E. Clarke (Eds.), *Developing Grounded Theory: The Second Generation* (pp. 194-234). Walnut Creek, California: Left Coast Press.
Cole, R. J., Oliver, A., & Robinson, J. (2013). Regenerative design, socio-ecological systems and co-evolution. *Building Research & Information, 41*(2), 237-247.
Coleman, D. C. (2010). *Big ecology: The emergence of ecosystem science*. University of California Press.
Collins, H., & Evans, R. (2008). *Rethinking expertise*. University of Chicago Press.
Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of The Royal Society, 368*, 4023-4038.
Cragin, M. H., Heidorn, P. B., Palmer, C. L., & Smith, L. C. (2007). *An educational program on data curation*. Paper presented at the American Library Association Conference, Science and Technology Section, Washington, D.C.. Available at https://www.ideals.illinois.edu/handle/2142/3493
Cragin, M. H. (2009). Scientific data collections: Use in scholarly communication and implications for data curation. (PhD), University of Illinois at Urbana-Champaign.
Crall, A. W., Newman, G. J., Jarnevich, C. S., Stohlgren, T. J., Waller, D. M., & Graham, J.

(2010). Improving and integrating data on invasive species collected by citizen scientists. *Biological Invasions, 12*(10), 3419-3428.
Creswell, J. (2007). *Qualitative inquiry and research design: Choosing among five approaches*: SAGE Publications.
Cushing, J. B., Kaplan, N. E., Laney, C., Mallett, J., Ramsey, K., Vanderbilt, K., … LeRoy, C. (2008). *Integrating ecological data: Notes from the Grasslands ANPP Data Integration Project.* Paper presented at the Environmental Information Management Conference 2008 Albuquerque, New Mexico, September 10--11, 2008.
Czarniawska, B. (2005). Karl Weick: Concepts, style and reflection. *The Sociological Review, 53*(s1), 267-278.
Darch, P. T., Borgman, C. L., Traweek, S., Cummings, R. L., Wallis, J. C., & Sands, A. E. (2015). What lies beneath?: Knowledge infrastructures in the subseafloor biosphere and beyond. *International Journal on Digital Libraries, 16*(1), 61-77.
Datacite (nd). Available at http://schema.datacite.org.
DataOne (nd). Data observation network for earth. Available at https://www.dataone.org.
Denzin, N. K., & Lincoln, Y. (2011). *Qualitative research*. Los Angeles: Sage.
DigiTool (nd). Digital resource manager. Available at https://knowledge.exlibrisgroup.com/DigiTool
Dougherty, D. J. (1987). *New products in old organizations: The myth of the better mousetrap in search of the beaten path* (PhD), MIT.
Dropbox (nd). Available at https://www.dropbox.com/about.
Duerr, R. E., Parsons, M. A., Marquis, M., Dichtl, R., & Mullins, T. (2004). *Challenges in long-term data stewardship.* Paper presented at the Twenty-First IEEE Conference on Mass Storage Systems and Technologies (MSST). University of Maryland.
Duerr, R. E., Downs, R. R., Tilmes, C., Barkstrom, B., Lenhardt, W. C., Glassy, J., … Slaughter, P. (2011). On the utility of identification schemes for digital earth science data: An assessment and recommendations. *Earth Science Informatics, 4*, 139-160.
Durrance, J. C., Walker, D., Souden, M., & Fisher, K. E. (2006). The role of community-based, problem-centered information intermediaries in local problem solving. *Proceedings of the American Society for Information Science and Technology, 43*(1), 1-17.
EcoPrairie Collection (nd). EcoPrairie project collection. Available at https://dspace.library.colostate.edu/handle/10217/87239.
EcoRiver Outreach (nd). EcoRiver outreach website. Available at http://www.experienceemiquon.com.
Edwards, P. N. (2003). Infrastructure and modernity: Force, time, and social organization in the history of sociotechnical systems. In T. J. Misa, P. Brey & A. Feedberg (Eds.), *Modernity and Technology* (pp. 185-225). Cambridge: MIT Press.
Edwards, P. N., Jackson, S., Bowker, G., & Knobel, C. P. (2007). Understanding infrastructure: dynamics, tensions, and design. Ann Arbor: National Science Foundation.
Edwards, P. N., Bowker, G. C., Jackson, S. J., & Williams, R. (2009). Introduction: an agenda for infrastructure studies. *Journal of the Association for Information Systems, 10*(5), 364-374.
Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., … Calvert, S. (2013). Knowledge infrastructures: Intellectual frameworks and research challenges. Ann Arbor: Report of a workshop sponsored by the National Science Foundation and the Sloan Foundation, Deep Blue.

Ehn, P. (1989). *Work-Oriented design of computer artifacts*. Hillsdale, New Jersey: Lawrence Erlbaum Assoc.
Ellison, A. M., Osterweil, L. J., Clarke, L., Hadley, J. L., Wise, A., Boose, E., . . . Kuzeja, P. (2006). Analytic webs support the synthesis of ecological data sets. *Ecology, 87*(6), 1345-1358.
Emerson, R. M., Fretz, R. I. & Shaw, L. L. (2011). *Writing ethnographic fieldnotes*. University of Chicago Press.
Erway, R. (2013). *Starting the conversation: University-Wide research data nanagement policy*: ERIC.
EZID (nd). Identifiers made easy. Available at http://ezid.cdlib.org.
Falzon, M.-A. (2009). *Multi-Sited Ethnography: Theory, praxis and locality in contemporary research*: Ashgate Publishing, Ltd.
Farrall, S. (2006). What is qualitative longitudinal research. *Papers in Social Research Methods Qualitative Series No, 11*.
Ferguson, A. R., Nielson, J. L., Cragin, M. H., Bandrowski, A. E., & Martone, M. E. (2014). Big data from small data: data-sharing in the 'long tail' of neuroscience. *Nature neuroscience, 17*(11), 1442-1447.
Fischer, G. (2013). A conceptual framework for computer-supported collaborative learning at work. In S. Goggins, I. Jahnke & V. Wulf (Eds.), *CSCL at Work*. Heidelberg: Springer.
Fisher and Durrance (2003)
Fisher, K. E., Landry, C. F., & Naumer, C. (2006). Social spaces, casual interactions, meaningful exchanges: 'information ground' characteristics based on the college student experience. *Information Research, 12*(2)
Fisher, K. E., & Durrance, J. C. (2003). Information communities. In K. Christen & D. Levinson (Eds.), *The Encyclopedia of Community: From the Village to the Virtual World* (pp. 657-660). Thousand Oaks, CA: SAGE Reference.
Floyd, C., Mehl, W.-M., Reisin, F.-M., Schmidt, G., & Wolf, G. (1989). Out of Scandinavia: Alternative approaches to software design and system development. *Human-Computer Interaction, 4*(4), 253-350.
Franklin, J. F., Bledsoe, C. S., & Callahan, J. T. (1990). Contributions of the Long-Term Ecological Research Program. *BioScience, 40*(7), 509-523.
Galison, P. (1997). *Image and logic: A material culture of microphysics*. University of Chicago Press, Chicago.
Geertz, C. (1973). *The interpretation of cultures: Selected essays* (Vol. 5019): Basic books.
Geospatial Centroid Center (nd). Geospatial Centroid Center. Available at http://gis.colostate.edu.
Gjefsen, M. D., & Fisher, E. (2014). From ethnography to engagement: The lab as a site of intervention. *Science as Culture, 23*(3), 419-431.
Godstein, J. C., Mayernik, M. S., Ramapriyan, H. K. (2017). Identifiers for Earth Science Data Sets: Where we have been and where we need to go. Data Science Journal, 16(23). Doi: http://doi.org/10.5334/dsg-2017-023.
Golley, F. B. (1993). *A history of the ecosystem concept in ecology: More than the sum of the parts*: Yale University Press.
Goodwin, C. (1995). Seeing in depth. *Social Studies of Science, 25*(2), 237-274.
Gosz, J. R., Waide, R. B., & Magnuson, J. J. (2010). Twenty-eight years of the US-LTER program: Experience, results, and research questions. In F. Muller & C. Bassler (Eds.),

Long-Term Ecological Research (Vol. Long-Term Ecological Research, Between Theory and Application, pp. 59-74): Springer.
Greenbaum, J., & Kyng, M. (Eds.). (1991). *Design at Work: Cooperative Design of Computer Systems*. Hillsdale, New Jersey: Lawrence Erbaum Associates.
Halskov, K., & Hansen, N. B. (2015). The diversity of participatory design research practice at PDC 2002–2012. *International Journal of Human-Computer Studies, 74*, 81-92.
Hanseth, O., Monteiro, E., & Hatling, M. (1996). Developing information infrastructure standards: the tension between standardisation and flexibility. *Science, Technology & Human Values, 21*(4), 407-426.
Hartter, J., Ryan, S. J., MacKenzie, C. A., Parker, J. N., & Strasser, C. A. (2013). Spatially explicit data: Stewardship and ethical challenges in science. *PLoS Biology, 11*(9), e1001634.
Harvey, P., Jensen, C. B., & Morita, A. (2017). *Infrastructures and social complexity: A companion*: Routledge.
Hays, R. G. (1980). *State science in Illinois: The scientific surveys, 1850-1978*: Board of Natural Resources and Conservation of the Illinois Institute of Natural Resources.
Heath, D. (2007). Bodies, antibodies and modest interventions. In K. Asdal, B. Brenna & I. Moser (Eds.), *Technoscience: The Politics of Interventions* (pp. 135-156).
Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends, 57*(2), 280-299.
Henderson, A., & Kyng, M. (1991). There's no place like home: Continuing design in use. In J. Greenbaum & M. Kyng (Eds.), *Design at Work: Cooperative Design of Computer Systems*. Hillsdale, New Jersey: Lawrence Erbau Associates.
Henshaw, D. L., Sheldon, W. M., Remillard, S. M., & Kotwica, K. (2006). *ClimDB/hydroDB: a web harvester and data warehouse approach to building a cross-site climate and hydrology database.* Paper presented at the Proceedings of the 7th International Conference on HydroScience and Engineering, Philadelphia, USA.Henshaw, 2006
Higgins, S. (2012). The lifecycle of data management. In G. Pryor (Ed.), *Managing Research Data* (pp. 17-45): Facet Publishing.
Hine, C. (2007). Multi-sited ethnography as a middle range methodology for contemporary STS. *Science, Technology & Human Values, 32*(6), 652-671.
Hobbie, J. E. (2003). Scientific accomplishments of the long term ecological research program: An introduction. *BioScience, 53*(1), 17-20.
Holdren, J.P. (2013). Increasing access to the results of federally funded scientific research, Memorandum for the Heads of Executive Departments and Agencies, U.S. Office of Science and Technology, February 22, 2013. Available at http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.
Hruby, J. M., Manley, D. K., Stoltz, R. E., Webb, E. K., & Woodard, J. B. (2011). The evolution of federally funded research and development centers. *Pubic Interest Report.*
Humphrey, C. (2006). e-Science and the life cycle of research. In CEOS (Ed.), *Data Life Cycle Models and Concepts*. University of Alberta, Canada. Available at http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc.
Hunter, N. C., Legg, K., & Oehlerts, B. (2010). Two librarians, an archivist, and 13,000 images: Collaborating to build a digital collection. *The Library Quarterly, 80*(1), 81-103.

Hunter, N., Rettig, P., & Liu, S. (2008). *Metadata matters: Developing a university best practices model.* Paper presented at the ELUNA 2008, 30 July - 1 August 2009, Long Beach, CA.

IBP Reports (nd). International Biological Program reports in the digital collections of Colorado. Available at https://dspace.library.colostate.edu/handle/10217/100252.

ICPSR (2013). Sustaining domain repositories for digital data: A call for change from an interdisciplinary working group of domain repositories. Available at http://www.icpsr.umich.edu/icpsrweb/ICPSR/support/announcements/2013/09/sustaining-domain-repositories-for.

ICPSR (nd). Data stewardshp. Available at https://www.icpsr.umich.edu/icpsrweb/content/about/data-stewardship.html.

INHS (nd). Illinois natural history survey field stations. Available at http://wwx.inhs.illinois.edu/fieldstations.

Jackson, S. J., & Barbrow, S. (2015). *Standards and/as innovation: Protocols, creativity, and interactive systems development in ecology.* Paper presented at the Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.

Jackson, S. J., Edwards, P. N., Bowker, G. C., & Knobel, C. P. (2007). Understanding infrastructure: History, heuristics and cyberinfrastructure policy. *First Monday*, *12*(6).

Jackson, S. J., Ribes, D., Bowker, G.C. & Buyuktur, A. (2010). Exploring collaborative rhythm: temporal flow and alignment in collaborative scientific work. Paper presented at the Proceedings of the iConference (Feb 3-6, 2010), Urbana-Champaign, IL.

Jasanoff, S., Martello, M., Haas, P. M., & Rochlin, G. (Eds.). (2004). *Earthly Politics: Local and Global in Environmental Governance*: The MIT Press.

Jensen, C. B. (2007). Infrastructural fractals: Revisiting the micro—macro distinction in social theory. Environment and Planning D: *Society and Space, 25*(5), 832-850.

Jensen, T. E. (2012). Intervention by invitation: New concerns and new versions of the user in STS. *Science Studies, 25*(1), 13-36.

Jensen, C. B., & Morita, A. (2016). Infrastructures as ontological experiments. *Ethnos*, 1-12.

Johnston, L. (Ed.). (2016). *Curating Research Data: Practical Strategies for Your Digital Repository (Vol. 1)* Chicago, Illinois: Association of College and Research Libraries.

Junk, W. J., Bayley, P. B., & Sparks, R. E. (1989). The flood pulse concept in river-floodplain systems. *Canadian Special Publication of Fisheries and Aquatic Sciences, 106*(1), 110-127.

Kansa, E. C., Kansa, S. W., & Arbuckle, B. (2014). Publishing and pushing: Mixing models for communicating research data in archaeology. *International Journal of Digital Curation, 9*(1), 57-70.

Kaplan, N. E., Vanderbilt, K., Zeman, L., Cushing, J.B., Laney, C., Mallett, J. … Muldavin, E. (2007). A team approach to data synthesis: The playbook for creating a centralized, dynamic, and sustainable ANPP database. Poster. LTER Regional Symposium, Albuquerque, New Mexico, July 11-12, 2008. Available at http://hdl.handle.net/10217/85125.

Kaplan, N. E., Draper, D. C., Paschal, D. B., Moore, J. C., Baker, K. S., & Swauger, S. (2014a). *Data curation issues in transitioning a field science collection of long-term research data and artefacts from a local repository to an institutional repository* (poster). Presented at the International Digital Curation Conference (IDCC), San Francisco, CA.

Kaplan, N. E., Baker, K. S., Draper, D. C., & Swauger, S. (2014b). *Packaging, transforming, and migrating data from a scientific research project to an institutional repository: The SGS LTER collection*. Digital Collections of Colorado, Colorado State University, Fort Collins, Colorado. Available at http://hdl.handle.net/10217/87239.

Karasti, H. (2014). *Infrastructuring in participatory design.* Paper presented at the Proceedings of the 13th Participatory Design Conference: Research Papers-Volume 1.

Karasti, H., & Baker, K. S. (2004). *Infrastructuring for the long-term: Ecological information management.* Paper presented at the Proceedings of the 37th Annual Hawaii International Conference on System Sciences.

Karasti, H., & Syrjänen, A.-L. (2004). *Artful infrastructuring in two cases of community PD.* Paper presented at the Proceedings of the eighth conference on Participatory design: Artful integration: interweaving media, materials and practices-Volume 1.

Karasti, H., Baker, K. S., & Halkola, E. (2006). Enriching the notion of data curation in e-science: Data managing and information infrastructuring in the Long Term Ecological Research (LTER) network. *Computer Supported Cooperative Work, 15*(4), 321-358.

Karasti, H., Baker, K. S., & Millerand, F. (2010). Infrastructure time: Long-term matters in collaborative development. *Computer Supported Cooperative Work (CSCW), 19*(3-4), 377-415.

Kingsland, S. E. (2010). The role of place in the history of ecology. In I. Billick & M. V. Price (Eds.), *The Ecology of Place: Contributions of Place-Based Research to Ecological Understanding*. Chicago: University of Chicago Press.

Kingsland, S. E. (2005). *The evolution of American ecology 1890-2000*. Baltimore: The Johns Hopkins University Press.

Kirchner, T. B., Chinn, H., Henshaw, D., & Porter, J. (1995). Documentation standards for data exchange. In R. Ingersoll & J. Brunt (Eds.), *Proceedings of the 1994 LTER Data Management Workshop*. Seattle, Washington: Long-Term Ecological Research Network Office, University of Washington.

Knorr-Cetina, K. D. (1992). The couch, the cathedral, and the laboratory: On the relationship between experiment and laboratory in science. In A. Pickering (Ed.), *Science as Practice and Culture* (pp. 113-137). Chicago: University of Chicago Press.

Koch, C., & Chan, P. (2013). *Projecting an infrastructure-shaping a community.* Paper presented at the Proceedings EPOC 2013 Conference.

Kowalczyk, S., & Shankar, K. (2011). Data sharing in the sciences. *Annual Review of Information Science and Technology, 45*(1), 247-294.

Kozlowski, S. W., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological science in the public interest, 7*(3), 77-124.

Laney, C. M., Baker, K. S., Peters, D. P. C., & Ramsey, K. W. (2013). Recommendations for Data Accessibility. In D. P. C. Peters, C. M. Laney, A. E. Lugo, S. L. Collins, C. T. Driscoll, P. M. Goffman, … J. Yao (Eds.), *Long-Term Trends in Ecological Systems: A Basis for Understanding Responses to Global Change* (pp. 216-225): United States Department of Agriculture, Agricultral Research Service.

Latzko-Toth, G., Bonneau, C., & Millette, M. (2017). Small data, thick data: Thickening strategies for trace-based social media research. *The SAGE handbook of social media research methods*, 199-214.

Lauenroth, W., & Burke, I. (2008). *Ecology of the Shortgrass Steppe: A long-term perspective (Long-Term Ecological Research Network)*: Oxford University Press.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*: Cambridge University Press.
Lee, H.-L. (2000). What is a collection? *Journal of the American Society for Information Science, 51*(12), 1106-1113.
Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues, 2*(4), 34-46.
Lewis, M. W., & Dehler, G. E. (2000). Learning through paradox: A pedagogical strategy for exploring contradictions and complexity. *Journal of Management Education, 24*, 708-725.
Likens, G. E., Cole, J. J., Kolasa, J., McAninch, J. B., McDonnell, M. J., Parker, G. G., & Strayer, D. L. (1987). Status and future of ecosystems science - Cary Conference, 28 April to 1 May 1985. *Occasional Publication of The Institute of Ecosystem Studies, 3*, 1-23.
Lindley, S. E., Thieme, A., Taylor, A. S., Vlachokyriakos, V., Regan, T., & Sweeney, D. (2017). Surfacing Small Worlds through Data-In-Place. *Computer Supported Cooperative Work (CSCW)*, 1-29.
Long, J. C., Cunningham, F. C., & Braithwaite, J. (2013). Bridges, brokers and boundary spanners in collaborative networks: A systematic review. *BMC Health Services Research, 13*(1), 1.
Lord, P., & Macdonald, A. (2003). Data curation for e-science in the UK: An audit to establish requirements for future curation and provision, e-Science Curation Report, : JCSR.
LTAR (nd). USDA/ARS/Long Term Agroecological Research network. Available at http://www.ars.usda.gov/research/programs/programs.htm?np_code=211&docid=22480.
LTER DataBits (nd). LTER DataBits, an Information Management Newsletter. Available at http://databits.lternet.edu.
LTER Data Portal (nd). Long term ecological research data portal. Available at https://lternet.edu/node/83507
LTER IM (2009). Review criteria for LTER information management. Available at http://im.lternet.edu/im_requirements/im_review_criteria.
LTER NIS (nd). Network information system website. Available at https://nis.lternet.edu:8443/display/NISnew.
Lyle, J. T. (1996). *Regenerative design for sustainable development*: John Wiley & Sons.
Lynch, C. A., & Lippincott, J. K. (2005). Institutional repository deployment in the United States as of early 2005. *D-Lib Magazine, 11*(9), 1-11.
Lynch, C. (2008). How do your data grow? *Nature, 455*.
MacMullin, S. E., & Taylor, R. S. (1984). Problem dimensions and information traits. *The Information Society, 3*(1), 91-111.
Marcus, G. E. (2007). Collaborative imaginaries. *Taiwan Journal of Anthropology, 5*(1), 1-17.
Marcus, G. E. (1995). Ethnography in/of the world system: The emergence of multi-sited ethnography. *Annual review of anthropology*, 95-117.
Mauz, I., Peltola, T., Granjou, C., Van Bommel, S., & Buijs, A. (2012). How scientific visions matter: Insights from three long-term socio-ecological research (LTSER) platforms under construction in Europe. *Environmental Science & Policy, 19*, 90-99.
Mayernik, M. S. (2016). Research Data and Metadata Curation as Institutional Issues. *Journal of the Association for Information Science and Technology, 67*(4), 973-993. doi: 10.1002/asi.23425.

Mayernik, M. S., Choudhury, G. S., DiLauro, T., Metsger, E., Pralie, B., Rippin, M., & Duerr, R. (2012). The Data Conservancy Instance: Infrastructure and Organizational Services for Research Data Curation. *D-Lib Magazine, 18*(9/10).

Mayernik, M. S., Daniels, M. D., Dattore, R. E., Davis, E. R., Ginger, K., Kelly, K. M., … Wright, M. J. (2012). Data citations within NCAR/UCP. Boulder, CO: National Center for Atmospheric Research, NCAR Technical Note: NCAR/TN-492+STR

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.* Paper presented at the Theory and Practice of Digital Libraries--TPDL 2013, Selected Workshops, Valletta, Malta, September 22-26, 2013.

Mayernik, M. S., Thompson, C. A., Williams, V., Allard, S., Palmer, C. L., & Tenopir, C. (2015). Enriching education with exemplars in practice: Iterative development of data curation internships. *International Journal of Digital Curation, 10*(1), 123-134.

Mayernik, M. S., Wallis, J. C., & Borgman, C. L. (2013). Unearthing the infrastructure: Humans and sensors in field-based scientific research. *Computer Supported Cooperative Work (CSCW), 22*(1), 65-101.

McNiff, J. (1988). *Action research: Principles and practice*. New York: Routledge.

Michener, W. K. (1986). *Research data management in the ecological sciences*: University of South Carolina Press.

Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B., & Stafford, S. G. (1997). Nongeospatial Metadata For The Ecological Sciences. *Ecological Applications, 7*(1), 330-342.

Michener, W. K., Miller, A., & Nottrott, R. (1990). *Long-Term Ecological Research Network Core Data Set Catalog*: Long-Term Ecological Research Network.

Michener, W. K., Porter, J., Servilla, M., & Vanderbilt, K. (2011). Long term ecological research and information management. *Ecological Informatics, 6*, 13-24.

Millerand, F., & Baker, K. S. (2010). Who are the users? Who are the developers? Webs of users and developers in the development process of a technical standard. *Information Systems Journal, 20*(2), 137-161.

Millerand, F., & Bowker, G. C. (2009). Metadata standards: Trajectories and enactment in the life of an ontology. In M. Lampland & S. L. Star (Eds.), *Standards and Their Stories. How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life* (pp. 149-165): Cornell University Press.

Millerand, F., Ribes, D., Baker, K., & Bowker, G. C. (2013). Making an Issue out of a standard: Storytelling practices in a scientific community. *Science, Technology & Human Values, 38*(1), 7-43.

Mills, H. B. (1958). *A century of biological research* (Vol. 27): Ayer Publishing.

Monteiro, E., Pollock, N., Hanseth, O., & Williams, R. (2013). From artefacts to infrastructures. *Computer Supported Cooperative Work (CSCW), 22(3)*, 575-607.

Myers, M. (1999). Investigating information systems with ethnographic research. *Communications of the AIS, 2*(4es).

NAP (2016). *The future of atmospheric chemistry research: Remembering yesterday, understanding today, anticipating tomorrow*: National Academies of Sciences.

Nardi, B. A., & O'Day, V. L. (2000). *Information ecologies: Using technology with heart*: MIT Press.

NAS (2009). *On being a scientist: A guide to responsible conduct of research.* Committee on Science, Engineering, and Public Policy.
NCAR (2012). National Center for Atmospheric Research, 2012 Annual Report. Boulder, Co. Available at http://www.nar.ucar.edu/2012/lar/ncar.html.
NCAR (2014a). 2014 NCAR Annual Report. Boulder, CO. Available at https://nar.ucar.edu/2014/ncar/2014-ncar-annual-report.
NCAR (2014b) NCAR Strategic Plan 2014-2019. Boulder, CO. Available at https://ncar.ucar.edu/directorate/documents/ncar-strategic-plan-2014-2019
NGRREC (nd). National Great Rivers Research and Education Center. Available at http://www.ngrrec.org.
NRC (1995). Colleges of Agriculture at the Land Grant Universities: A Profile. National Research Council, National Academies Press. Available at http://www.nap.edu/download/4980.
NRC (2003). NEON: Addressing the nation's environmental challenges. In N. R. C. Committee on the National Ecological Observatory Network (Ed.). National Academies Press.
NRC (2007). Strategic Guidance for the National Science Foundation's Support of the Atmospheric Sciences: National Research Council.
NRC (2015). Preparing the workforce for digital curation: Committee on future career opportunities and educational requirements for digital curation; Board on Research Data and Information; Policy and Global Affairs.
NSB (2005). Long-lived digital data collections enabling research and education in the 21st Century, *National Sciences Board*.
NSTC (2009). Harnessing the Power of Digital Data for Science and Society: Report of the Interagency Working Group on Digital Data. Available at http://www.nitrd.gov/About/Harnessing_Power_Web.pdf.
Orlikowski, W. J. (2002). Knowing in practice: Enacting a collective capability in distributed organizing. *Organization Science*, 13(3), 249–273.
Orr, J. E. (1996). *Talking about machines: An ethnography of a modern job.* Ithaca, **NY** Cornell University Press.
Palmer, C. L. (2004). Thematic research collections. In S. Schreibman, R. Siemens & J. Unsworth (Eds.), *A Companion to Digital Humanities* (Vol. 21, pp. 348-365). Oxford: Blackwell.
Palmer, C. L., Cragin, M. H., Heidorn, P. B., & Smith, L. C. (2007). *Data curation for the long tail of science: The case of environmental sciences.* Paper presented at the Third International Digital Curation Conference, Washington, DC.
Palmer, C. L., Thomer, A. K., Baker, K. S., Wickett, K. M., Hendrix, C., Rodman, A., . . . Fouke, B. W. (2017). Site-Based Data Curation Based on Hot Spring Geobiology. *PloS One*. DOI: 10.1371/journal.pone.0172090.
Palmer, C. L., Thompson, C. A., Mayernik, M. S., Williams, V., Kelly, K., & Allard, S. (2014). *Data curation education in research centers: Formative evaluation findings from 2012-2013 cohorts.* Presented at the 9th International Digital Curation Conference, February 24-27, 2014, San Francisco, CA.
Palmer, C. L., Mayernik, M. S., Weber, N., Baker, K. S., Kelly, K., Marlino, M. R., & Thompson, C. A. (2013a). *Research problems in data curation: Outcomes from the data curation education in research centers program.* Paper presented at the American Geophysical Union Fall Meeting Abstracts San Francisco, CA.

Palmer, C., Renear, A., and Cragin, M. (2008). Purposeful curation: Research and education for a future with working data. Proceedings of the 4th International Digital Curation Conference.

Palmer, C. L., Weber, N. M., Munoz, T., & Renear, A. H. (2013b). Foundations of data curation: The pedagogy and practice of "purposeful work" with research data. *Archive Journal*(3).

Palmer, C. L., Zavalina, O. L., & Fenlon, K. (2010). Beyond size and search: Building contextual mass in digital aggregations for scholarly use. *Proceedings of the American Society for Information Science and Technology, 47*(1), 1-10.

Pampel, H., & Dallmeier-Tiessen, S. (Eds.). (2014). *Open Research Data: From Vision to Practice*: Springer.

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., . . . Dierolf, U. (2013). Making research data repositories visible: the re3data. org registry. *PloS One, 8*(11). doi: doi:10.1371/journal.pone.0078080

Parsons, M. A., Godøy, Ø., LeDrew, E., De Bruin, T. F., Danis, B., Tomlinson, S., & Carlson, D. (2011). A conceptual framework for managing very diverse data for complex, interdisciplinary science. *Journal of Information Science, 37*(6), 555-569.

Pasquetto, I. V., Sands, A. E., Darch, P. T., & Borgman, C. L. (2016). Open data in scientific settings: From policy to practice. Paper presented at the Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.

PASTA QC (nd). Provenance aware synthesis tracking architecture data package quality checks. Available at http://im.lternet.edu/sites/im.lternet.edu/files/Data_package_quality_checks_Report_july2012.pdf.

Peters, D. P. C., Laney, C. M., Lugo, A. E., Collins, S. L., Driscoll, C. T., Goffman, P. M., … Yao, J. (2013). Long-term trends in ecological systems: A basis for understanding responses to global change (Vol. Technical Bulletin Number 1931. Available at http://www.ars.usda.gov/is/np/LongTermTrends/LongTermTrendsIntro.htm).

Peters, D. P.C., Loescher, H. W., SanClements, M. D., & Havstad, K. M. (2014). Taking the pulse of a continent: Expanding site-based research infrastructure for regional-to continental-scale ecology. *Ecosphere, 5*(3), art29.

Pickett, S. T. A., & Cadenasso, M. L. (2008). Linking ecological and built components of urban mosaics: An open cycle of ecological design. *Journal of Ecology*, *96*, 8-12.

Pipek, V., & Wulf, V. (2009). Infrastructuring: Toward an integrated perspective on the design and use of information technology. *Journal of the Association for Information Systems, 10*(5), 447.

Pomerantz, J. (2008). Digital (library services) and (digital library) services. *Journal of Digital Information, 9*(2).

Porter, J. H., & Callahan, J. T. (1994). Circumventing a dilemma: Historical approaches to data sharing in ecological research. In W. K. Michener, J. W. Brunt & S. G. Stafford (Eds.), *Environmental Information Management and Analysis: Ecosystem to Global Scales* (pp. 193-202): Taylor & Francis.

Porter, J. H. (2010). A brief history of data sharing in the US Long Term Ecological Research Network. *Bulletin of the Ecological Society of America, 91*(1), 14-20.

Pryor, G. (2012). Why manage research data? In G. Pryor (Ed.), *Managing Research Data* (pp. 1-16): Facet Publishing.

Ray, J. M. (2014). *Research data management: Practical strategies for information professionals*: Purdue University Press.
RDA (nd). Research Data Alliance, Foundation and Terminology Working Group. Available at https://rd-alliance.org/groups/data-foundation-and-terminology-wg.html.
Re3data (nd). Data repository registry. Available at http://re3data.org.
Reason, P., & Bradbury, H. (2008). *The SAGE handbook of action research, 2nd Edition*: Sage.
Renear, A. H., Sacchi, S., & Wickett, K. M. (2010). Definitions of dataset in the scientific and technical literature. *Proceedings of the American Society for Information Science and Technology, 47*(1), 1-4.
Rettig, P. J., Liu, S., Hunter, N., & Level, A. V. (2008). Developing a metadata best practices model: The experience of the Colorado State University libraries. *Journal of Library Metadata, 8(4)*.
Ribes, D., & Baker, K. (2007). Modes of social science engagement in community infrastructure design *Communities and Technologies 2007* (pp. 107-130): Springer.
Ribes, D., & Lee, C. P. (2010). Sociotechnical studies of cyberinfrastructure and e-research: Current themes and future trajectories. *Computer Supported Cooperative Work, 19*, 231-244.
RIN (2009). Patterns of information use and exchange: Case studies of researchers in the life sciences. In R. Williams & G. Pryor (Eds.), *Research Information Network and the British Library*.
Rolland, K. H., Hepsø, V., & Monteiro, E. (2006). *Conceptualizing common information spaces across heterogeneous contexts: mutable mobiles and side-effects of integration.* Paper presented at the Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work.
Roth, W.-M., & Bowen, G. M. (2001). Of disciplined minds and disciplined bodies: On becoming an ecologist. *Qualitative Sociology, 24*(4), 459-481.
Rücknagel, J., Vierkant, P., Ulrich, R., Kloska, G., Schnepf, E., Fichtmüller, D., … Kirchhoff, A. (2015). Metadata schema for the description of research data repositories (re3data.org). re3data.org: GFZ Germans Research Center for Geosciences.
Saad-Sulonen, J. (2013). Multiple Participations In L. Horelli (Ed.), *New Approaches to Urban Planning* (pp. 111-130). Helsinki, Finland: Aalto University.
San Gil, I., Hutchison, V., Frame, M., & Palanisamy, G. (2010). Metadata activities in biology. *Journal of Library Metadata, 10*(2-3), 99-118.
Sandvig, C. (2013). The internet as infrastructure. *The Oxford Handbook of Internet Studies*, 86-108.
Schmidt, K. (2010). *"Keep up the good work!": The concept of "work" in CSCW*. Proceedings of COOP 2010. Computer Supported Cooperative Work. Computer Supported Cooperative Work, DOI 10.1007/978-1-84996-211-7_15.
Schmidt, K. (2011a). The concept of 'work'in CSCW. *Computer Supported Cooperative Work (CSCW), 20*(4-5), 341-401.
Schmidt, K. (2011b). *Cooperative Work and Coordinative Practices, Contributions to the Conceptual Foundations of Computer-Supported Cooperative Work (CSCW)*: Springer.
Schmidt, K., & Bannon, L. (1992). Taking CSCW seriously, Supporting articulation work. *Computer Supported Cooperative Work, 1*(1-2), 7-40.
Schmidt, K., & Bansler, J. (2016). Computational artifacts: Interactive and collaborative computing as an integral feature of work practice. Paper presented at the COOP 2016:

Proceedings of the 12th International Conference on the Design of Cooperative Systems, 23-27 May 2016, Trento, Italy.
Schon, D. (1983). *The reflective practitioner: How professionals think in action*: Basic Books.
Schuler, D., & Namioka, A. (1993). *Participatory design: Principles and practices*: CRC.
Schumpeter, J. A. (1939). *Business cycles*. New York: McGraw Hill. (from NSTC, 2009)
Servilla, M., & Brunt, J. (2011). The LTER Network Information System: Improving data quality and synthesis through community collaboration. *AGU Fall Meeting Abstracts, 1*, 1598.
Servilla, M., Brunt, J., Costa, D., McGann, J., & Waide, R. (2016). The contribution and reuse of LTER data in the Provenance Aware Synthesis Tracking Architecture (PASTA) data repository. *Ecological Informatics*. doi:10.1016/j.ecoinf.2016.07.003
Servilla, M., Brunt, J., San Gil, I., & Costa, D. (2006). PASTA: A network-level architecture design for automating the creation of synthetic products in the LTER Network. *LTER Information Management Databits Newsletter Spring 2006.*
Servilla, M., Costa, D., Laney, C., San Gil, I., & Brunt, J. (2008). The EcoTrends web portal: An architecture for data discovery and exploration. Paper presented at the Proceedings of the Environmental Information Management Conference.
Shaon, A., Giaretta, D., Crompton, S., Conway, E., Matthews, B., Marelli, F., … Guarino, R. (2012). *Towards a Long-term preservation infrastructure for earth science data.* Paper presented at the Proceedings of the 9th Internernational Conference on Digital Preservation (iPres' 2012).
Siggelkow, N. (2007). Persuasion with case studies. *Academy of Management Journal, 50*(1), 20-24.
Simonsen, J., & Robertson, T. (Eds.). (2013). *Handbook of Participatory Design*. New York: Routledge.
Sinclair, R. A. (1983). Long Term Ecological Research Data Management Workshop, November 22-23, 1982. In R. A. Sinclair (Ed.). Urbana-Champaign, Illinois: Illinois State Water Survey, Illinois Department of Energy and Natural Resources.
Singh, S. J., Haberl, H., Chertow, M., Mirtl, M., & Schmidt, M. (Eds.). (2012). *Long Term Socio-Ecological Research: Studies in Society: Nature Interactions Across Spatial and Temporal Scales* (2). Springer.
Sparks, R. E. (1992). The Upper Mississippi river restoration of aquatic ecosystems: Science, technology, and public policy. In National Academy Press (Series Ed.) (pp. 406-411). Washington, D.C.: In National Academy Press. Available at http://www.nap.edu/catalog.php?record_id=1807.
Sparks, R. E. (1995). Need for ecosystem management of large rivers and their floodplains. *BioScience, 45*(3), 168-182.
Sparks, R. E. (2010). Forty years of science and management on the Upper Mississippi River: An analysis of the past and a view of the future. *Hydrobiologia, 640*(1), 3-15.
Sparks, R. E. (2016). Controversy and science at the meeting of great rivers. *Elsah History, Summer 2016*(109 & 110), 7-19.
Sparks, R. E., Bayley, P. B., Kohler, S. L., & Osborne, L. L. (1990). Disturbance and recovery of large floodplain rivers. *Environmental management, 14*(5), 699-709. SPIC (nd). Scripps Pelagic Invertebrate Collection. Available at https://scripps.ucsd.edu/collections/pi.
Spinuzzi, C. (2005). The methodology of participatory design. *Technical Communication, 52*(2), 163-174.

Spinuzzi, C. (2013). *Topsight: A guide to studying, diagnosing, and fixing information flow in organizations*.

Stafford, S. G., Kaplan, N. E., & Bennett, C. W. (2002). *Through the Looking Glass: What do we see, What have we learned, What can we share? Information Management at the Shortgrass Steppe Long Term Ecological Research Site.* Paper presented at the International Institute of Informatics and Systematics (SCI2002), July 14-18, 2002, Orlando, Florida.

Stake, R. E. (1995). *The art of case study research*. Thousand Oaks, CA: Sage Publications.

Star, S. L. (1999). The Ethnography of Infrastructure. *American Behavioral Scientist, 43*(3), 377-391.

Star, S. L. (2000). *It's infrastructure all the way down (keynote address)*. Paper presented at the Proceedings of the fifth ACM conference on Digital libraries, San Antonio, Texas, USA.

Star, S. L., & Bowker, G. C. (2002). How to Infrastructure. In L. A. Lievrouw & S. Livingstone (Eds.), *Handbook of New Media: Social Shaping and Consequences of ICTs* (pp. 151-162). London: Sage Publications.

Star, S. L., & Ruhleder, K. (1994). *Steps towards an ecology of infrastructure: Complex problems in design and access for large-scale collaborative systems.* Paper presented at the Proceedings of the 1994 ACM conference on Computer supported cooperative work.

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information systems research, 7*(1), 111-134.

Star, S. L., & Strauss, A. (1999). Layers of silence, arenas of voice: The ecology of visible and invisible work. *Computer Supported Cooperative Work (CSCW), 8*, 9-30.

Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R. R., Duerr, R., … Clark, T. (2015). Achieving human and machine accessibility of cited data in scholarly publications. Available at https://doi.org/10.7717/peerj-cs.1.

Steinhart, G., Chen, E., Arguillas, F., Dietrich, D., & Kramer, S. (2012). Prepared to plan? A snapshot of researcher readiness to address data management planning requirements. *Journal of eScience Librarianship, 1*(2), 1.

Strasser, C., Cook, R., Michener, W., & Budden, A. (2012). Primer on data management: What you always wanted to know: Dataone.org.

Strauss, A. (1978). A Social World Perspective. *Studies in Symbolic Interaction, 1*, 119-128.

Strauss, A., & Corbin, J. (1998). Basics of qualitative research: Procedures and techniques for developing grounded theory: Thousand Oaks, CA: Sage.

Suchman, L. (1987). *Plans and situated actions: The problem of human-machine communication*: Cambridge University Press.

Suchman, L. (2001). Building bridges: Practice-based ethnographies of contemporary technology. *Anthropological perspectives on technology*, 163-177.

Suchman, L. (2002a). Located accountabilities in technology production. *Scandinavian Journal of Information Systems, 14*(2), 7.

Suchman, L. (2002b). Practice-based design of information systems: Notes from the hyperdeveloped world. *The Information Society, 18*, 139-144.

Susman, G. I., & Evered, R. D. (1978). An assessment of the scientific merits of action research. *Administrative Science Quarterly*, 582-603.

Swan, A., & Brown, S. (2008). The skills, role and career structures of data scientists and curators: An assessment of current practice and future needs. London: JISC. Available at http://eprints.soton.ac.uk/266675/.

Swanson, F. J., & Sparks, R. E. (1990). Long-Term Ecological Research and the invisible place. *BioScience, 40*(7), 502-508.
Taylor, R. S. (1991). Information use environments. In B. Dervin (Ed.), *Progress in Communication Sciences* (Vol. 10, pp. 217-225). Norwood, NJ: Ablex.
Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PloS One, 6*(6), e21101. doi: 10.1371/journal.pone.0021101
Tenopir, C., Birch, B., & Allard, S. (2012). *Academic libraries and research data services: Current practices and plans for the future; an ACRL White Paper*: Association of College and Research Libraries, a division of the American Library Association.
Thompson, C. A. (2015). Building data expertise into research institutions: Preliminary results. *Proceedings of the Association for Information Science and Technology, 52*(1), 1-5.
Thompson, C. A., Mayernik, M. S., Palmer, C. L., Allard, S., & Tenopir, C. (2015). LIS programs and data centers: Integrating expertise. *iConference 2015 Proceedings*.
TNC (2006). Key attributes and indicators for Illinois River conservation targets at the Nature Conservancy's Emiquon Preserve: The Nature Conservancy of Illinois.
Treloar, A., Choudhury, G. S., & Michener, W. (2012). Contrasting national research data strategies: Australia and the USA. In G. Pryor (Ed.), *Managing Research Data* (pp. 173-203): Facet Publishing.
TRS (2012). Science as an open enterprise, open data for open science. The Royal Society Science Policy Centre. Available at http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf.
Twidale, M. B., & Floyd, I. (2008). *Infrastructures from the bottom-up and the top-down: Can they meet in the middle?* Paper presented at the Proceedings of the Tenth Anniversary Conference on Participatory Design 2008.
U.S. Libraries (nd). Available at http://www.library.gov.
Van den Hoven, J. (1999). Information resource management: Stewards of data. *Information Systems Management, 16*(1), 88-90.
Vanderbilt, K., Cushing, J., Gao, J., Kaplan, N., Kruger, J., Leroy, C., … Zeman, L. (2009). Data integration challenges: An example from the International Long-Term Ecological Research Network (ILTER). *Ecological Circuit, 2*, 12.
Van House, N. A. (2004). Science and technology studies and information studies. *Annual Review of Information Science and Technology, 38*, 3-86.
Varvel, V. E. J., Palmer, C. L., Chao, T., & Sacchi, S. (2011). Report from the research data workforce summit: Sponsored by the Data Conservancy. Champaign, IL, USA. Available at http://hdl.handle.net/2142/25830.
Venturi, R. (1966). *Complexity and contradiction in architecture* (Vol. 1): The Museum of Modern Art, New York.
Vertesi, J. (2014). Seamful spaces: Heterogeneous infrastructures in interaction. *Science, Technology & Human Values, 39*, 264-284.
Walk, J. W., Baker, K. S., & Sparks, R. E. (2016). Data stewardship workshop report, March 2016. Therkildsen Field Station at Emiquon The Nature Conservancy and University of Illinois. Available at http://hdl.handle.net/2142/94785.
Walk, J. W., Lemke, M. J., Lemke, A. M., & Sparks, R. E. (in prep). Emiquon: Introduction to a large-scale floodplain restoration. *Hydrobiologia*.

Wallis, J. C. (2012). *The distribution of data management responsibility within scientific research groups.* (PhD), University of California, Los Angeles, Los Angeles. Available at http://www.escholarship.org/uc/item/46d896fm.

Wallis, J. C., Borgman, C. L., Mayernik, M. S., & Pepe, A. (2008). Moving archival practices upstream: An exploration of the life cycle of ecological sensing data in collaborative field research. *International Journal of Digital Curation, 3*(1), 114-126.

Wang, Y., Kaplan, N., Newman, G., & Scarpino, R. (2015). CitSci. org: A new model for managing, documenting, and sharing citizen science data. *PLoS Biol, 13*(10), e1002280.

Wayback Machine (nd). The wayback machine. Available at https://archive.org/web/.

Weick, K. E. (1995). What theory is not, theorizing is. *Administrative Science Quarterly, 40*(3), 385-390.

Weick, K. E. (2016). Constrained comprehending: The experience of organizational inquiry. *Administrative Science Quarterly, 61*(3), 333-346.

Whyte, A. (2012). Emerging infrastructure and services for research data management and curation in the UK and Europe. In G. Pryor (Ed.), *Managing Research Data* (pp. 205-234): Facet Publishing.

Wickett, K. M., Isaac, A., Fenlon, K., Doerr, M., Meghini, C., Palmer, C. L., & Jett, J. (2013). Modeling cultural collections for digital aggregation and exchange environments: Center for informatics research in science and scholarship, CIRSS Technical Report 201310-1.

Wilbanks, J. (2011). Openness as infrastructure. *Journal of ChemInformatics, 3*, 36.

Wilkins-Diehr, N. (2007). Special issue: science gateways—common community interfaces to grid resources. *Concurrency and Computation: Practice and Experience, 19*(6), 743-749.

Willig, M. R., & Walker, L. R. (Eds.). (2016). *Long-Term Ecological Research: Changing the Nature of Scientists*. New York: Oxford University Press.

Wyatt, S., & Balmer, B. (2007). Home on the range: What and where is the middle in science and technology studies? *Science, Technology, & Human Values, 32*(6), 619-626.

Yeo, G. (2012). Bringing things together: Aggregate records in a digital age. *Archivaria, 74*.

Zimmerman, A. (2007). Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal of Digital Libraries, 7*, 5-16.

Zimmerman, A., & Finholt, T. A. (2007). Growing an infrastructure: The role of gateway organizations in cultivating new communities of users. Paper presented at the Proceedings of the 2007 international ACM conference on Supporting Group Work.